УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Лука Перовић

ПРОБЛЕМ ДВА УЗОРКА У ОБРАДИ ТЕКСТА

мастер рад

Београд, 2024.

Ментор:

др Бојана Милошевић, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Марко Обрадовић, доцент
Универзитет у Београду, Математички факултет

др Марија Цупарић, доцент
Универзитет у Београду, Математички факултет

Датум одбране: _____

Желим да изразим своју искрену захвалност свом ментору, др Бојани Милошевић, за њено руководство, увид и континуирану помоћ током овог истраживања. Захвалност упућујем и др Ани Колуби и Луизи Контопиорпис на пруженим подацима. Такође сам захвалан својим колегама др Марку Обрадовићу и др Марији Цупарић, на њиховим вредним повратним информацијама и сарадњи. На крају, захвалио бих се породици и пријатељима на подршци коју су ми пружили.

Наслов мастер рада: Проблем два узорка у обради текста

Резиме: У раду описујемо пробабилистичке тематске моделе, класу модела који откривају латентне структуре у групама докумената. Даћемо математичку основу над којом је изграђен модел латентне алокације Дрихлеа грађен, пре него што опишемо сам модел. Дискутоваћемо процес евалуације модела овог типа, описујући модеран приступ кохерентности тема, са фокусом на C_V кохеретност, најсавременију у овој класи мера. При том ћемо дати детаљно математичко објашњење и мотивацију иза ове мере.

У остатку рада ћемо се фокусирати на меру присутности, која служи да процени значај екстерно задате теме у групи докумената, користећи притом параметре пробабилистичких тематских модела. Описаћемо и тест два узорка заснован на овој статистици.

На крају ћемо дати два примера примене латентне алокације Дирихлеа и мере присутности. Прво ћемо анализирати теме које се појављују у онлајн рецензијама женске одеће. Меру присутности ћемо да применимо тако што меримо количину позитивног сентимента у рецензијама. Анализираћемо резултате мере присутности и валидирати их преко оцена рецензената. За други пример ћемо оценити тематски модел над апстрактима радова са CMStatistics конференције. Меру присутност користимо да анализирамо утицај пројекта на тематску структуру радова на конференцији.[1]

Кључне речи: статистичка обрада текста, латентне структуре, пробабилистички тематски модели, латентна алокација Дирихлеа, Дирихлеова расподела, кохерентност тема, присутност, тест два узорка докумената, анализа сентимента

[1]Кодове коришћене у раду је могуће добити уз захтев. Контакт мејл: lukarajkin@gmail.com

Садржај

1 Увод **1**
1.1 Статистичка обрада текста 1
1.2 Тематско моделирање 2

2 Математичке основе **5**
2.1 Дирихлеова расподела 5
2.2 Нормализована тачкаста узајамна информација 9

3 Тематски модели **13**
3.1 Структура података, матрица докумената и термина 13
3.2 Латентна алокација Дирихлеа 15

4 Мере квалитета тематских модела **18**
4.1 Тест речи уљеза 19
4.2 Аутоматске мере кохерентности теме 20

5 Присутност **23**
5.1 Тест присутности два узорка 25

6 Примена **28**
6.1 Рецензије женске одеће 28
6.2 Књиге апстраката конференције CMStatistics 35

7 Закључак **39**

Библиографија **41**

Глава 1

Увод

1.1 Статистичка обрада текста

Статистика се користи у обради текста још од средине 20. века. Џорџ Зиф (енг. George Zipf), амерички лингвиста и филолог је још у првој половини 20. века анализирао статистичке особине језика, дајући емпиријски закон о вези броја појављивања речи и њеном рангу појављивања (в. [30]).

Недетерминистичка природа језика је интуитивно довела до тога да се језик моделира случајним процесом, где се процес генерисања језика описује расподелом вероватноћа над вокабуларом речи, условљен претходним речима у тексту.

Средином 20. века је амерички математичар Клод Шенон (енг. Claude Shannon) изнео своју теорију информације (в. [24]) и дефинисао неодређеност као општу меру насумичности категоричке случајне променљиве. У чланку је дао доњу и горњу границу за неодређеност енглеског језика, дајући статистичку основу анализи текста. Тада се појављује један од првобитних модела језика: $n-$грам модел (в. [24]), који вероватноћу појављивања речи у тексту моделира у односу на претходних $n-1$ речи, наивно процењујући вероватноћу на основу фреквенције појављивања дате комбинације n речи. У седамдесетим годинама почињу да се користе скривени Марковљеви модели, који претпостављају да иза речи у тексту стоји неопсервабилни Марковљев ланац. Ови модели постају популарни и ефективни у препознавању говора и обележавању речи по врстама (енг. Part of Speech Tagging, в. [16]). Инспирисани теоријом информације, деветесетих година истраживачи у IBM-у популаризују статистичко превођење, третирајући проблем превођења са јед-

ног језика на други као проблем декодирања (в. [7]).

Иако се почетком 21. века појављују први неурални модели језика (в. [1]), биће им потребно дуго време да покажу напредак у односу на раније поменуте статистичке моделе. Гугл је, примера ради, у свом онлајн преводиоцу, тек 2016. године прешао на неурални модел (в. [29]).

Последњих година, дубоке неуронске мреже засноване на архитектури трансформера (в. [26]) показују изванредне перформансе у разним пољима обраде текста. Велики језички модели (енг. Large Language Models, в.[8]) на основу огромне количине текста успевају да уоче јако комплексне структуре и везе у језику.

Ипак, ови модели, са милијардама параметара, имају великуману интерпретабилности. Неретко су названи црне кутије, јер се посматрају само у виду улаза и излаза, без икаквог разумевања њиховог интерног процеса.

Ми ћемо се зато у овом раду бавити једном другом класом модела, такозваним тематским моделима. Развијени већински почетком 21. века, дају другачији приступ структури језика. Супротно од дубоких неуронских мрежа, ови модели дају јасну пробабилистичку слику, са фокусом на интерпретабилношћу. У другом делу рада ћемо се бавити статистиком заснованом на параметрима модела, ослањајући се на јасно значење и интерпретацију ових параметара (в. [4, 21]).

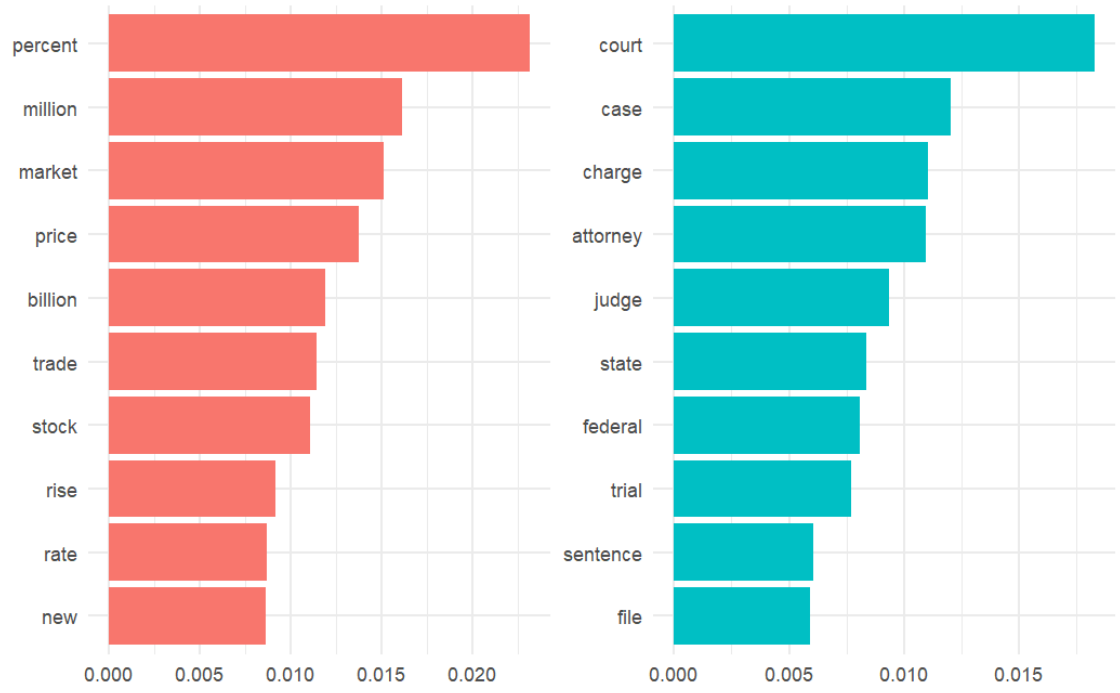
1.2 Тематско моделирање

Тематско моделирање представља грану обраде природног језика и ненадгелданог машинског учења, која за циљ има да пронађе скривене тематске структуре у великој количини докумената.

Претпоставка тематског модела јесте да текст, односно речи које се налазе у документу, потиче из тема, које су неопсервабилне, скривене (латентне) променљиве. Саме теме су представљене терминима (речима) које их сачињавају, и ничим другим. Основни задатак тематских модела јесте да на математички смислен начин открију ову скривену структуру. Резултат оваквог модела јесу додељене теме за сваки документ и додељене речи за сваку тему.

Пробабилистичко тематско моделирање даље претпоставља да се документи могу описати као расподеле над темама, а теме као расподеле над терминима. Оваква претпоставка има за природну предност да индивиду-

ални документ не припада једној јединој теми, већ јасно приказује колико је која тема присутна у документу, у виду своје расподеле вероватноћа над темама. Такође даје подлогу за налажење кључних речи унутар тема, у виду рангирања речи по вероватноћи унутар теме (в. [2]).



Слика 1.1: Пример две издвојене теме пронађене над чланцима њујоршких новина AssociatedPress.

На слици 1.1 можемо да видимо практичан пример изгледа тема у пробабилистичком тематском моделу. Прва тема има везе са берзом, док се друга бави правосудним системом. Приказано је најважнијих 10 термина за обе теме, рангирани по вероватноћи појављивања у теми. Коришћен модел јесте LDA (Латентна алокација Дирихлеа, в. [4]), најраспрострањенији пробабилистички тематски модел и модел на који ћемо се фокусирати у овом раду.

Ова класа модела је нашла примену у разним сферама. Првобитно је развијена ради ефикасног организовања архива, са идејом да се одабере тема из области интересовања и да се брзо нађу документи у којима је та тема присутна (в. [4]). При читању одређеног документа, на пример новинског чланка, можда би корисника занимали документи који су тематски слични. У пољима онлајн претраге, постоји потреба да се нађу документи у којима

су честе теме у којима су битни термини које је корисник унео. У пољима надгледаног машинског и статистичког учења можемо да користимо пробабилистичке тематске моделе као облик претпроцесуирања у форми смањења димензионалности текстуалних података. Уместо чистог текста, који у општем случају представља низ речи произвољне дужине можемо документ да представимо преко његове оцењене расподеле над темама, којих је коначно много (в. [4, 2]).

При анализи анкета, честа су питања отвореног типа. Анализа оваквих одговора иначе захтева велику количину ручног рада. Тематско моделирање се овде може користити у сврху препознавања честих образаца у одговорима (в. [6]). Примене су бројне, и нису ограничене само на математичке и информационе. Једна од првих примене ове класе модела је заправо била у популационој генетици (в. [21]), где су ови модели независно развијени. Иако смо представили тематске моделе у контексту примене над текстом, са математичке стране текст није ништа више него вектор категоричких променљивих произвољне дужине. Ако генетски код особе посматрамо као документ, теме заправимо представљају честе шаблоне у генетском коду. Ови шаблони могу да дају увид у порекло особе и да разазнају различите етничке средине наших предака. У сфери текста, сваки документ настаје из мешавине тема. У сфери генетике, генетски код особе потиче из различитих прапостојбина предака.

Глава 2

Математичке основе

Пре него што опишемо моделе, увешћемо неке математичке појмове, битне за разумевање остатка рада.

2.1 Дирихлеова расподела

Дирихлеова расподела (в. [20]) представља вишедимензионо уопштење бета расподеле. Као таква, обухвата фамилију апсолутно непрекидних вишедимензионих расподела. Често се користи као априорна расподела у Бајесовој стастистици, због чињенице да је конјугована расподела мултиномне расподеле. Ово и представља мотивацију иза њеног коришћења у тематском моделирању.

Дефиниција 2.1.1 *Дирихлеова расподела над $k-1$-димензионим симплексом*[1] *параметризована вектором позитивних реалних бројева $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$. Густина расподеле Дирихлеове расподеле дата је са:*

$$f(\mathbf{x}; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

[1] n-димензиони симплекс је представљен формулом:

$$\left\{ \mathbf{x} \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} x_i = 1 \text{ и } x_i \geq 0 \ \forall i \in \{0, 1, \ldots, n\} \right\}$$

ūge je

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}.$$

Пошто је носач расподеле $k-1$-димензони симплекс, случајне промењљиве са Дирихлеовом расподелом могу да се користе да моделирају било какве тежине чији је збир 1, а посебно и расподеле вероватноћа мултиномних расподела. Од сада па надаље ћемо да задржимо ову интерпретацију и причаћемо о реализованим вредностима Дирихлеове расподеле као расподели вероватноћи.

Особине и интуиција иза расподеле

У овом одељку ћемо да изложимо неке особине Дирихлеове расподеле, са циљем да дамо бољи увид у природу расподеле.
Прво ћемо одредити маргиналне расподеле и очекивања.
Пошто су промењљиве ограничене везом $\sum_{i=1}^{k} x_i = 1$, замењујемо x_k са $x_k = 1 - \sum_{i=1}^{k-1} x_i$. Вредност маргиналне густине је онда:

$$f(x_1; \alpha_1, \ldots, \alpha_k) = \frac{1}{\mathrm{B}(\alpha)} x_1^{\alpha_1 - 1} \int_0^{1-x_1} \cdots \int_0^{1-\sum_{i=1}^{k-2} x_i} \left(1 - \sum_{i=1}^{k-1} x_i\right)^{\alpha_k - 1} \prod_{i=2}^{k-1} x_i^{\alpha_i - 1} dx_{k-1} \cdots dx_2.$$

Унутрашњи интеграл по x_{k-1} можемо израчунати сменом $t_{k-1} = \frac{x_{k-1}}{1-\sum_{i=1}^{k-2} x_i}$. Ова смена своди унутрашњи интеграл на облик бета расподеле. Понављајући овај поступак долазимо до

$$f(x_1; \alpha_1, \alpha_0) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0 - \alpha_1)} x_1^{\alpha_1 - 1} (1 - x_1)^{\alpha_0 - \alpha_1 - 1}$$

где је $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Резултат је густина бета расподеле, односно

$$X_1 \sim \mathrm{Beta}(\alpha_1, \alpha_0 - \alpha_1).$$

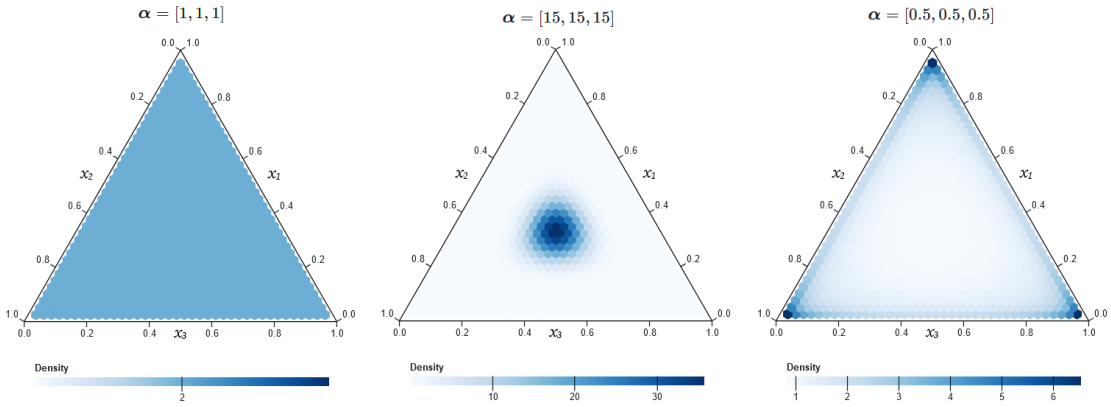
Аналогно следи да је маргинална расподела било које компоненте Дирихлеове расподеле бета расподела, односно

$$X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i).$$

На основу овога лако налазимо да је очекивање расподеле

$$E[X_i] = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}.$$

Што је параметар α_i већи, то је очекивање i-те компоненте веће, тј. у просеку ће i-та вероватноћа генерисане мултиномне расподеле бити већа.
Из очекивања се ипак не види цела слика, јер скалирање свих параметара не утиче на очекивање, али знатно мења саму расподелу. Приказаћемо неколико примера густине Дирихлеове расподеле за $k = 3$ ради могућности визуелизације. На слици 2.1 видимо пример три симетричне Дирихлеове расподеле.



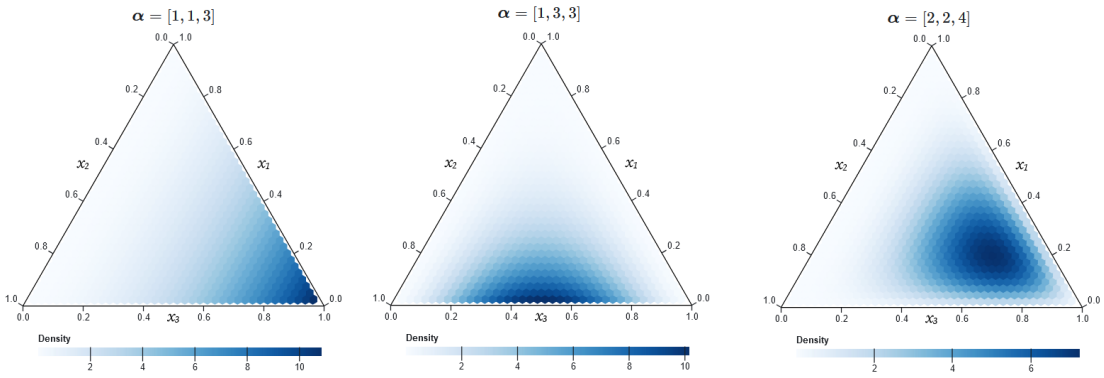
Слика 2.1: Пример три симетричне Дирихлеове расподеле.
Слика направљена уз помоћ [25].

У сваком случају је очекивање једнако $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. У првом примеру видимо униформну расподелу над 2-симплексом. У другом примеру је вредност параметара > 1, што има за последицу да расподела фаворизује униформније мултиномне расподеле. У трећем примеру, када је вредност параметра < 1, видимо прорећену расподелу, која фаворизује екстремне мултиномне расподеле, где је један догађај знатно вероватнији од друга два.

Нама ће од користи највише бити симетричне Дирихлеове расподеле, јер при коришћењу Дирихлеове расподеле као априорне расподеле, без неких додатних претпоставки, природно је третирати све случајеве једнако. Видимо из примера на слици 2.1 да параметар α контролише колико је расподела прорећена.

Конјугованост са мултиномном расподелом

Посматрајмо мултиномну расподелу са k могућих исхода и параметром вероватноћа исхода $p = (p_1, p_2, \ldots, p_k)$. Претпоставимо да имамо узорак обима



Слика 2.2: Пример неколико асиметричних Дирихлеових расподела.
Слика је направљена уз помоћ [25].

N и нек n_i представљају фреквенције догађаја у узорку, тако да је $\sum_{i=1}^{k} n_i = N$.

Функција веродостојности је тада:

$$P(\mathbf{n} \mid p) = \frac{N!}{n_1! n_2! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

где је $\mathbf{n} = (n_1, n_2, \dots, n_k)$.

Претпоставимо Дирихлеову априорну расподелу са параметром $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$. Апостериорна расподела параметра p је тада пропорционална производу веродостојности и априорне густине:

$$P(p \mid \mathbf{n}, \alpha) \propto P(\mathbf{n} \mid p) \cdot P(p \mid \alpha)$$

Убацивањем у једначину добијамо:

$$P(p \mid \mathbf{n}, \alpha) \propto \prod_{i=1}^{k} p_i^{n_i + \alpha_i - 1}$$

Препознајемо Дирихлеову расподелу са параметром $\alpha' = (\alpha'_1, \alpha'_2, \dots, \alpha'_k)$, где:

$$\alpha'_i = \alpha_i + n_i$$

Дакле, апостериорна расподела параметра p на основу узорка $\mathbf{n}$ је:

$$p \mid \mathbf{n}, \alpha \sim \mathrm{Dir}(\alpha_1 + n_1, \alpha_2 + n_2, \dots, \alpha_k + n_k)$$

Овим смо показали конјугованост датих расподела.
Када бисмо прошли очекивањем кроз горњи израз, добили бисмо тачкасте Бајесове оцене параметра p:

$$E[p_i \mid \mathbf{n}] = \frac{n_i + \alpha_i}{N + \sum_{j=1}^{k} \alpha_j}$$

Ради једноставно, посматрајмо случај симетричне Дирихлеове расподеле:

$$E[p_i \mid \mathbf{n}] = \frac{n_i + \alpha}{N + k\alpha}$$

Упоређујући са оценама максималне веродостојности, које представљају једноставне релативне фреквенције категорија, коришћење априорне Дирихлеове расподеле нам изглађује оцене вероватноћа тако што сваку приближава вредности $\frac{1}{k}$. Што је параметар α већи, ово изглађивање је јаче.

Ово има значајан пожељан ефекат код мултиномних расподела са великим бројем категорија. Када мултиномна расподела има велики број категорија, у зависности од величине узорка, је врло вероватно да се нека категорија не појави у узорку, иако има позитивну вероватноћу. У случају језичког моделирања, када категорије представљају речи, можемо да направимо велику грешку и претпоставимо да је појављивање одређене речи немогуће, што би, при било каквој будућој евалуацији модела, при појављивању те речи, дало веродостојност модела једнаку нули.

2.2 Нормализована тачкаста узајамна информација

У овом поглављу ћемо да уведемо неке мере асоцијације речи, које ће нам касније бити битне при евалуацији тематских модела.

Дефиниција 2.2.1 *Нека су X и Y две дискретне случајне величине. Нека су x и y могуће вредности случајних величина X и Y, редом. У [10] је дата следећа дефиниција.* ***Тачкаста узајамна информација*** *(енг. PMI - Pointwise Mutual Information) x и y представља логаритам односа вероватноће остваривања обе вредности и производа вероватноћа остваривања индивидуалних вредности, тј.*

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

где подразумевамо $P(x, y) = P(X = x, Y = y)$, $P(x) = P(X = x)$ *и* $P(y) = P(Y = y)$.

Напомена 2.2.1 *У теорији информација је стандардно да логаритми буду у основи 2, како би све било мерено у битовима, иако није од суштинске важности, те ћемо се ми, ради једноставности, држати природног логаритма.*

Ова мера пореди стварну вероватноћу остваривања обе вредности са вероватноћом остваривања обе вредности, под претпоставком да су догађаји $\{X = x\}$ и $\{Y = y\}$ независни. У контексту обраде језика, где случајне величоне представљају појављивања одређених речи, мери колико се две речи чешће јављају заједно у односу на случајно појављивање. Овај термин је уско везан са чешћим термином узајамне информације.

Дефиниција 2.2.2 ***Узајамна информација*** *(енг. MI - Mutual Information) две дискретне случајне величине представља очекивање тачкасте узајамне информације:*

$$MI(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} = \mathbb{E}[PMI(X, Y)].$$

Тачкаста узајамна информација две вредности је позитивна, уколико се оне чешће појављују заједно него што би било очекивано под претпоставком независности. Она је једнака 0, ако јесу независни, а негативна ако се ређе јављају заједно него што би било за очекивати под условом независности.
У случају да вредности потпуно искључују једна другу, вредност PMI постаје $-\infty$, те је ова мера неограничена с доње стране.
С горње стране, максимум се достиже када вредности повлаче једна другу, тј. када је $P(x) = P(y) = P(x, y)$. Тада је вредност тачкасте узајамне информације

$$\mathrm{PMI}(x, y) = -\log P(x) = -\log P(y) = -\log P(x, y).$$

На основу овога можемо да назначимо референтне вредности за PMI:

- $\mathrm{PMI}(x, y) = -\infty$ ако вредности искључују један другог;
- $\mathrm{PMI}(x, y) = 0$ ако су независни;

- $\mathrm{PMI}(x, y) = -\log P(x, y)$ ако повлаче једна другу.

Пошто и вредност горње границе за произвољне вредности x и y може да буде произвољно велика, закључујемо да је тачкаста узајамна информација неограничена и са горње стране, односно

$$\mathrm{PMI}(x, y) \in (-\infty, \infty).$$

Ипак, горња граница за специфичне вредности (претпоставимо даље да су у питању појављивања речи) нам дају увид у особине ове мере. Наиме, две речи које се стално појављују заједно имају већу тачкасту узајамну информацију што су ређе. Ово у многим случајевима није пожељно понашање.

Ова мера се, пре коришћења у евалуацији тематских модела, највише користила у аутоматском издвајању фраза. У том пољу се посматрају парови речи, са циљем да се чести парови речи сврстају у јединствене фразе. Тачкаста узајамна информација би се користила као мера по којој би се поредили и рангирали парови речи. У ову сврху, ако се две речи увек појављују заједно, нема смисла их казнити као пар само зато што су честе. Штавише, са емпиријске стране, можда би логичније било више рангирати честе парове који су увек заједно, због чињенице да имамо више статистичког доказа о њиховом понашању.

У циљу кориговања ових недостатака, а и у постизању интуитивнијеих вредности мере, развијене су многе модификације, а ми ћемо издвојити *Нормализовану тачкасту узајамну информацију* (енг. NPMI - Normalised Pointwise Mutual Information, в. [5]).

Дефиниција 2.2.3 ***Нормализована тачкаста узајамна информација*** *две вредности x и y случајних величина X и Y је једнака*

$$NPMI(x, y) = \frac{PMI(x, y)}{-\log P(x, y)} = \frac{\log \frac{P(x,y)}{P(x)P(y)}}{-\log P(x, y)}.$$

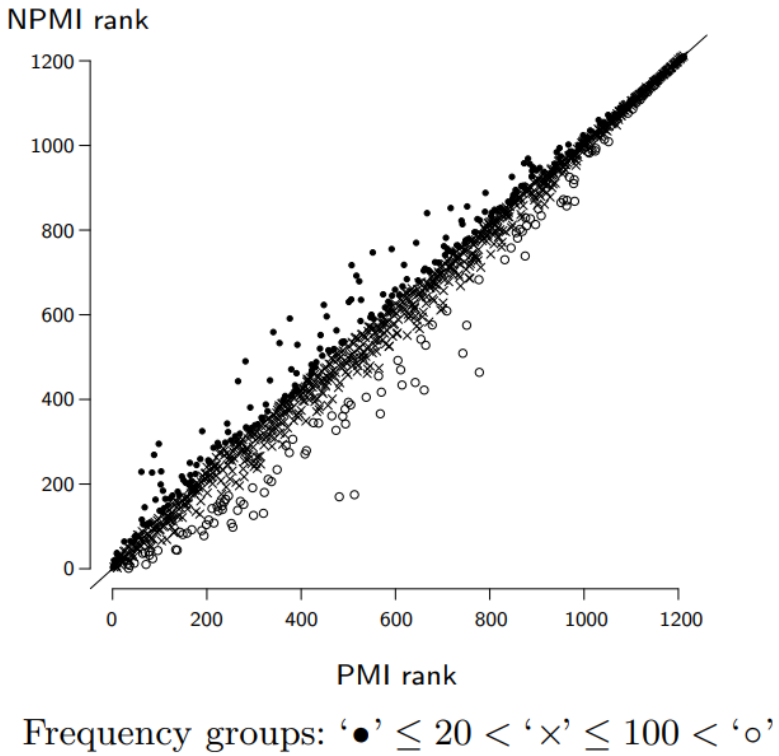
Већ смо видели да именилац у изразу представља горњу границу за PMI, што повлачи да је NPMI увек мањи или једнак 1.
Са друге стране, $-\log \frac{P(x,y)}{P(x)P(y)} \leq \log P(x, y)$, па се добија да је доња граница за NPMI једнака -1, односно

$$\mathrm{NPMI}(x, y) \in [-1, 1].$$

Закључујемо да су референтне вредности за NPMI:

- $\text{NPMI}(x, y) = -1$ ако догађаји искључују један другог;
- $\text{NPMI}(x, y) = 0$ ако су независни;
- $\text{NPMI}(x, y) = 1$ ако су еквивалентни.

Додаћемо да нормализовање тачкасте узајамне информације такође смањује њену фаворизацију према ретким терминима. На слици 2.3 видимо сличност



Слика 2.3: Упоређивање рангова парова речи по NPMI и PMI, груписаних до фреквенцији самих парова.
Слика узета из [5].

PMI и NPMI мера. Видимо и да је за високофреквентне парове ранг по NPMI знатно нижи, док је за нискофреквентне парове ранг по NPMI виши.

Глава 3

Тематски модели

У овом поглављу нам је главни циљ да дефинишемо модел Латентне алокације Дирихлеа (енг. LDA - Latent Dirichlet Allocation), један од првобитних тематских модела, који је до дан данас најпознатији у овој класи модела, који се подразумевано користи као репер за упоређивање сваког новонаправљеног пробабилистичког тематског модела, и који служи као основа на коју се већина тих модела надограђује.
Пре тога, објаснићемо структуру података са којом ови модели раде.

3.1 Структура података, матрица докумената и термина

Основна јединица података са којом радимо јесте документ. Један документ представља семантички повезану целину језика. Документ може да буде чланак из новина, глава из књиге, аптракт научног рада, итд.
Претпоставићемо да имамо скуп од M докумената који називамо *корпусом* $C = \{d_1, d_2, ..., d_M\}$.
Кад кажемо да радимо са скупом докумената, суптилно назначавамо да редослед докумената није битан. У неким тематским моделима се временски елемент документа узима у обзир (в. [3]), али нама неће бити од користи.
Сваки *документ* d_i представља низ речи коначне дужине $d_i = (w_1, w_2, ..., w_{N_i})$, где са N_i обележавамо дужину i-тог документа.
Свака *реч* јесте једна од V категорија. Скуп свих могућих категорија називамо *речником*, или *вокабуларом*. Величина вокабулара је тада V.

Матрица докумената и термина

Једна битна, у прву руку нелогична, претпоставка општих тематских модела, јесте да игноришу ред речи у документима. Ако имамо за циљ да моделујемо језик, ово је заиста необична и здраворазумски нетачна претпоставка. Ипак, као што смо навели у уводу, циљ ових модела није савршено моделовање језика, већ смислено и интерпретабилно извлачење теме из корпуса, и обележавање докумената у њема са тим темама. Касније, када будемо улазили у математичку дефиницију LDA модела, пробаћемо да дамо и објашњење зашто ова претпоставка може и да припомогне проналажењу тема. За сад ћемо узети ову претпоставку као начин појадностављења података.
Занемаривање редоследа речи има за последицу да се сваки документ може описати са бројем појављивања сваке од V речи у себи. Ово значи да наш скуп података можемо приказати као матрицу реда $M \times V$, где редови представљају различите документе, а колоне различите речи. У пресеку је број појављивања дате речи у датом документу. Ову матрицу називамо, уз мањак креативности, **матрицом докумената и термина**.
Она представља финални облик података са којим тематски модели оперишу.

	pant	comfortable	add	little	flare	typical	mom	pajama	look
text1	1	2	1	1	1	1	1	1	1
text2	0	1	0	0	0	0	0	0	0
text3	0	1	0	0	0	0	0	0	1
text4	1	0	0	0	0	0	0	0	1
text5	0	0	0	1	0	0	0	0	1
text6	0	0	0	0	0	0	0	0	0
text7	0	0	0	1	0	0	0	0	1
text8	0	0	0	0	0	0	0	0	0

Слика 3.1: Пример исечка матрице докумената и термина добијене над корпусом рецензија женске одеће.

3.2 Латентна алокација Дирихлеа

Латентна алокација Дирихлеа је пробабилистички тематски модел, што значи да претпоставља да сваки документ представља мешавину тема и да речи у документу потичу из тих тема.
Овај модел описује генеративни процес којим настају документи из датог корпуса.
Задржавајући ознаке из претходне секције, означићемо и број тема са K, унапред задати број, односно хиперпараметар модела.
Генеративни процес може да се објасни у неколико корака (в. [4]):

1. Узоркују се теме као расподеле над речима $\varphi_k \sim \mathrm{Dir}(\beta)$, где $k \in \{1, \ldots, K\}$ и $\mathrm{Dir}(\beta)$ је симетрична Дирихлеова расподела димензије V са скаларним параметром β.

2. Узоркују се расподеле тема за сваки документ $\theta_i \sim \mathrm{Dir}(\alpha)$, где $i \in \{1, \ldots, M\}$ и $\mathrm{Dir}(\alpha)$ је симетрична[1] Дирихлеова расподела димензије K са скаларним параметром α.

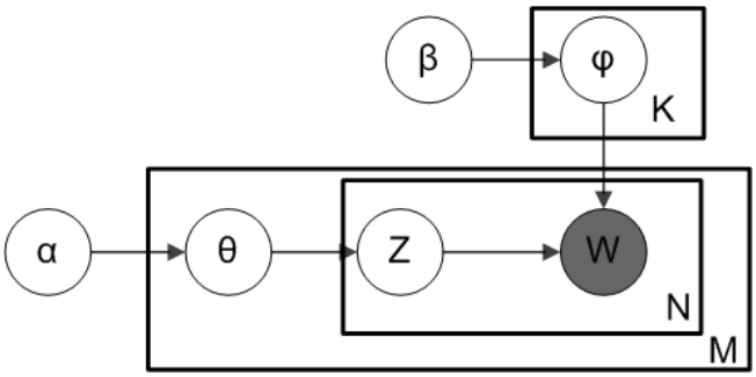
3. За сваку позицију речи i, j, где $i \in \{1, \ldots, M\}$ и $j \in \{1, \ldots, N_i\}$:

   a) Узоркује се тема $z_{i,j} \sim \mathrm{Cat}(\theta_i)$, где је Cat категоричка расподела са параметром који представља вектор вероватноћа категорија.

   b) Узоркује се реч $w_{i,j} \sim \mathrm{Cat}(\varphi_{z_{i,j}})$.

Као што можемо да видимо, основна идеја модела јесте да, након што генеришемо теме за дати документ, речи документа се онда генеришу као прост случајан узорак. Одавде следи да су речи међусобно независне под условом да знамо теме за документ, што повачи да је ред речи у документу небитан.
Ова условна независност речи, као што смо напоменули, не покушава да прецизно моделује језик, али сама по себи повлачи да теме морају да носе што више информација о речима. Модел тежи да сврста речи које се често налазе заједно у једну тему, јер су теме једини начин да искористи везу између речи уоквиру једног документа.

[1] Априорна Дирихлеова расподела за расподелу тема у документу не мора нужно да буде симетрична. У неким имплементацијама α може да буде вектор различитих вредности који се оптимизује у процесу тренинга модела. Ово има за предност да побољша генеративна својства модела, препознавајући које су теме чешће, а које ређе. Ми ћемо се држати једноставног случаја, где је α позната скаларна вредност.



Слика 3.2: Плочаста нотација генеративног процеса LDA модела. Стрелице означавају редослед. Бројеви у угловима плоча означавају независно понављање процеса дефинисаног унутар плоче.
Преузето из [23].

Приликом описивања генеративног[2] процеса који LDA претпоставља, успутно смо увели многе нове ознаке разних параметара. Сада ћемо их описати:

- β - Скаларни параметар априорне симетричне Дирихлеове расподеле за расподелу речи унутар тема. Хиперпараметар модела. Постављен често на мале вредности, пошто очекујемо да су теме сачињене од малог процента укуног вокабулара.

- φ_k, где $k \in \{1, \ldots, K\}$ - Категоричка расподела речи k-те теме. Скуп ових расподела по k можемо и посматрати матрично, где врсте представљају теме, а колоне речи. Ову матрицу ћемо једноставно да обележавамо са φ.

- α - Скаларни параметар априорне симетричне Дирихлеове расподеле за расподеле тема унутар докумената. Хиперпараметар модела. Вредност параметра α мења густину Дирихлеове расподелу, што утиче на то колико претпостављамо да су мешане теме по документима.

[2]Да би процес заиста био потпуно генеративан, потребно је да се дефинише процес генерисања броја речи сваког документа. У оригиналном раду је овај корак посебно издвојен, користећи Поасонову расподелу за моделирање дужине документа. Ипак, овај корак никако не утиче на оцењивање параметара који су нам важни, па се углавном изоставља ради једноставности.

- θ_i, где $i \in \{1, \ldots, M\}$ - Категоричка расподела тема i-тог документа. И ово можемо посматрати као матрицу θ, где су врсте документи, а колоне теме.

- $z_{i,j}$, где $i \in \{1, \ldots, M\}$ и $j \in \{1, \ldots, N_i\}$, тема j-те речи i-тог документа. Теме самих речи су латентне променљиве, нису обзервабилне и нису нам битне као излаз модела.

- $w_{i,j}$ - j-та реч i-тог документа. Једини опсервабилни параметри.

Глава 4

Мере квалитета тематских модела

Пробабилистичке језичке моделе, као и већину пробабилистичких модела, можемо евалуирати помоћу веродостојности модела. Једноставним речима, за сваки текст, при датим параметрима модела, посматрамо вероватноћу да модел генерише баш такав текст. Модел који боље описује структуру језика ће имати већу вероватноћу генерисања реалистичних текстова, и веродостојност модела ће бити виша.

Код пробабилистичких тематских модела приступ оригинално није био нимало другачији (в. [4]). Конкретно, због чињенице да строго комплекснији модел увек може боље да се прилагоди подацима, и кад није пригоднији, пракса је да се подаци поделе на тренинг и тест скуп. На тренинг скупу се оцене параметри модела, а на тест скупу се врши евалуација, тј. израчуна веродостојност тест скупа (енг. Held-out likelihood).

Такође се у литератури језичких модела, па и тематских, помиње збуњеност модела (енг. perplexity), мера из теорије информације, еквивалентна са веродостојношћу.

Ваљало би напоменути да рачунање веродостојности LDA модела није тривијално, због чињенице да се у изразу налазе несводљиви интеграли. Нећемо да улазимо превише у ове детаље, али бисмо истакли да је доста истраживања посвећено управо процени веродостојности (в. [27]).

Ипак, као што смо напоменули више пута у раду, тематски модели и не теже да савршено моделују структуру језика. Велики задатак тематских модела јесте да дају људски интерпретабилне теме. Препоручено је било да се при

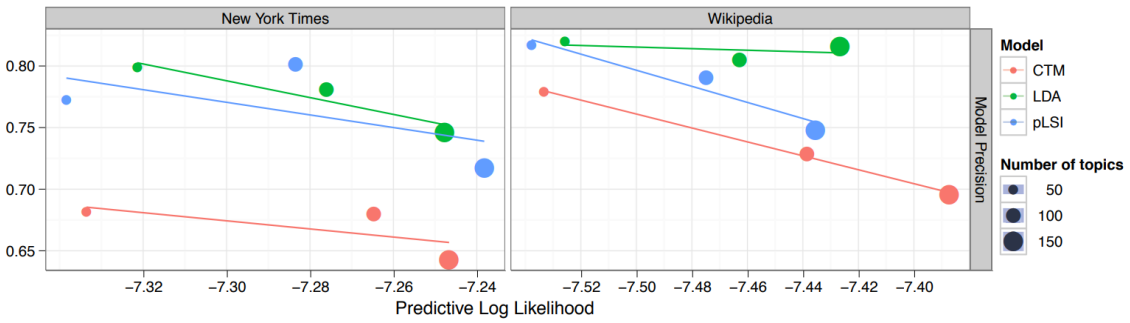
одабиру тематског модела, баци поглед на то како теме изгледају, и да се одабере онај модел који даје смисленије теме. Наиме, истраживачи би приказали теме, тј. најчешће речи сваке теме, слично као што смо ми у примеру у уводу, и одокативно проценили да ли теме могу да се интерпретирају.
Овај приступ, осим што је непрактичан при евалуацији великог броја модела, или модела са великим бројем тема, пати и од великог проблема субјективности.
У раду [9] под називом „Читање листића чаја: Како људи евалуирају тематске моделе", аутори вуку аналогију ка античкој традицији прорицања будућности на основу облика који листови чаја оставе на дну шоље. Напоменули су ипак, како је људска интерпретација битна за тематске моделе и предложили објективан начин за мерење *кохерентности* одређене теме, односно тога колико речи једне теме збирно дају смислену целину. Предложени метод је тест речи уљеза, који ћемо описати у наредном поглављу.

4.1 Тест речи уљеза

Тест речи уљеза (енг. Word Intrusion Test) је једноставан експеримент за мерење кохерентности теме:

1. За дату тему се издвоји 5 речи са највећим вероватноћама.
2. На тих 5 речи се насумично дода „уљез" - реч која има слабу тежину у датој теми, али је важна за неку другу тему модела.
3. Од испитаника се тражи да индентификује уљеза. Бележи се да ли је погодио.
4. Кораци 2 и 3 се понове велики број пута, са различитим испитаницима, како би се проценила вероватноћа одређивања уљеза.

Идеја је да је у смисленој теми лако наћи уљеза, док ће у бесмисленој теми погађање бити насумично.
Још један значајан, а доста неочекиван резултат изложен у раду [9] је био да су просечна кохерентност тема модела и његова веродостојност на тест скупу негативно корелисани.



Слика 4.1: Негативна корелисаност прецисности налажења уљеза и веродостојности модела. Узето из [9].

4.2 Аутоматске мере кохерентности теме

На крају рада [9] су аутори дали предлог да се због непрактичности људске евалуације усмери истраживање ка проналажењу аутоматске замене за мерење кохерентности теме.

Од тада су настале бројне мере. Укратко ћемо описати стандардан облик ових мера и описати две популарне мере кохерентности.

Слично као и у тесту уљеза, аутоматске мере кохерентности посматрају најважнијих N речи теме, где је N најчешће 10 или 5. Већина њих рачуна неку меру сличности између свих парова тих речи, коју после тога упросечи по паровима да би се добила коначна мера кохерентности.

Једна популарна мера кохерентности је **NPMI кохерентност**[1], уведена у [31]:

$$C_{\text{NPMI}} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{NPMI}(w_i, w_j)$$

где су w_i, $i \in \{1, 2, ..., N\}$, речи које имају највеће вероватноће у одређеној теми.

Претпоставка иза мере јесте да, ако речи припадају истој семантичкој теми, оне ће се често појављивати заједно, те ће им NPMI бити висок.

[1]Формално, морамо да дефинишемо случајне величине над којима се рачуна NPMI. Природно би било, на основу ознака, претпоставити да су то неке случајне величине чије вредности могу да буду речи w_i и w_j. Ипак, не можемо такве величине да дефинишемо а да ова мера задржи значење које желимо. Уместо тога, формално посматрамо случајне величине који су индикатори појављивања датих речи, односно $I\{\text{реч } w_i \text{ се појавила у тексту}\}$ и $I\{\text{реч } w_j \text{ се појавила у тексту}\}$. Кад пишемо $\text{NPMI}(w_i, w_j)$ мислимо на $\text{NPMI}(1, 1)$ за ове две случајне величине.

Аутоматске мере кохерентности су евалуиране по томе колику корелацију имају са људским мерилима, био то раније поменути тест уљеза или неки други метод рангирања тема.

У раду [22] су аутори, инспирасани разним, до тада откривеним, мерама кохерентности, дефинисали простор мера кохерентности (дефинисан са начином процене вероватноћа које фигуришу у рачуну сличности речи, саме мере сличности речи, начина агрегације, итд.). Претрагом по том простору, и обимном евалуацијом перформанси, испливала је нова мера као најбоља, $\mathbf{C}_V$ **кохерентност**.

C_V кохерентност исто користи NPMI, али не као коначну меру сличности речи. Процес је следећи:

1. За сваку од најважнијих N речи се прави вектор контекста $\vec{v_i}$, $i \in \{1, 2, .., N\}$. Елементи ових вектору представљају NPMI између речи $v_{ij} = \mathrm{NPMI}(w_i, w_j)$, $i, j \in \{1, 2, .., N\}$.

2. Рачуна се вектор контекста теме, као просек свих вектора контекста најважнијих N речи.
$$\vec{v_T} = \frac{1}{N} \sum_{i=1}^{N} \vec{v_i}$$

3. Рачуна се косинусно растојање између вектора контекста сваке од најважнијих N речи и вектора контекста теме.
$$\cos(\vec{v_i}, \vec{v_T}) = \frac{\vec{v_i} \cdot \vec{v_T}}{\|\vec{v_i}\|\|\vec{v_T}\|} = \frac{\sum_{j=1}^{N} v_{ij} v_{Tj}}{\sqrt{\sum_{j=1}^{N} v_{ij}^2} \sqrt{\sum_{j=1}^{N} v_{Tj}^2}}$$

4. Вредности добијене у претходном кораку се упросече по свакој од најважнијих N речи. Ово је финална мера кохерентности.
$$\mathrm{C}_V = \frac{1}{N} \sum_{i=1}^{N} \cos(\qquad Td \qquad \vec{} \qquad v \quad T \qquad Tj$$

пак да их буде више. Присутност предмета нам даје објективну меру екстерне теме и избегава проблеме субјективне претраге.

Посматрајмо корпус докумената као репрезентативан узорак популације. Нека је X случајна величина која представља присутност предмета по документу. Присутности докумената у корпусу представљају остварене вредности случајне величине X и обележавамо их са $x_1, x_2, ..., x_N$.

Дефиниција 5.0.3 ***Укупну присутност предмета*** *дефинишемо као очекивање дате случајне величине:*

$$\nu = E(X).$$

Укупна присутност је параметар који тежимо да оценимо. Стандардан начин оцене параметра очекивања јесте узорачка средина. Ипак, узорак присутности није опсервабилан, пошто стварни параметри LDA модела нису познати, већ су нам познате оцењене вредности тих параметара.

Оцену присутности термина у датом документу можемо да дефинишемо на следећи начин:

$$\hat{\nu}_{w_l}^{(i)} = \sum_{m=1}^{K} \hat{\theta}_{im} \hat{\varphi}_{ml}.$$

Оцену присутности предмета у датом документу исто добијамо заменом LDA параметара са њиховим оценама

$$\hat{x_i} = \hat{\nu}^{(i)} = \sum_{j=1}^{f} \mathbf{w}_j \hat{\nu}_{w_{l_j}} = \sum_{j=1}^{f} \sum_{m=1}^{K} \mathbf{w}_{l_j} \hat{\theta}_{i_m} \hat{\varphi}_{m_{l_j}}.$$

Коначно, оцена за укупну присутност се добија упросечавањем по свим доступним документима:

$$\hat{\nu} = \frac{1}{N} \sum_{i=1}^{N} \hat{\nu}^{(i)} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{f} \sum_{m=1}^{K} \mathbf{w}_{l_j} \hat{\theta}_{i_m} \hat{\varphi}_{m_{l_j}}.$$

5.1 Тест присутности два узорка

Конструисаћемо тест укупне присутности два независна[2] узорка.
При конструкцији теста, игнорисаћемо варијабилност унету приликом оцењивања параметара LDA модела, одосно, сматраћемо да су оцењене вредности

[2]Узорке не можемо формално заиста сматрати независнима, јер оба узорка зависе од оцењених параметара истог LDA модела. У наставку текста је дато објашњење.

$\hat{x}_1, \hat{x}_2, ..., \hat{x}_N$ стварне вредности. Ова претпоставка нам омогућава да апстракујемо математику иза рачунања присутности предмета у документима и посматрамо $x_1, x_2, ..., x_N$ као узорак из непознате расподеле.

Постојаност параметара LDA модела је истражена при различитим методама апроксимације, са закључком да је могуће конзистентно оцењивање LDA параметара. Ова чињеница нам даје основу за поменуто игнорисање варијабилности (в.[11, 28]).

Посматрајмо две потпопулације докумената. Означимо укупну присутност предмета у ове две популације са ν_1 и ν_2.
Тестирамо:

$$H_0 : \nu_1 = \nu_2$$
$$H_1 : \nu_1 \neq \nu_2.$$

Пошто игноришемо варијабилност у параметрима модела, можемо да искористимо класичну тест статистику за два узорка:

$$T = \frac{\hat{\nu}_1 - \hat{\nu}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}.$$

Због природне зависности у параметрима модела, не можемо да очекујемо нормалност мере присутности. Због тога користимо двоузорачки наивни бутстреп метод за процену расподеле тест статистике. Под нултом хипотезом претпостављамо да је иста просечна присутност у обе популације докумената. Ово можемо да, при бутстреп реузорковању, симулирамо тако што узоркујемо из центрираних узорака. Извршавање теста можемо описати у следећим корацима предложеним у [18]:

1. Израчунамо вредност тест статистике:

$$T = \frac{\hat{\nu}_1 - \hat{\nu}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

2. Реузоркујемо из центрираних подузорака $\{x_{11} - \hat{\nu}_1, \ldots, x_{1n_1} - \hat{\nu}_1\}$, $\{x_{21} - \hat{\nu}_2, \ldots, x_{2n_2} - \hat{\nu}_2\}$ како бисмо добили бутстреп оцене одговрајућих средњих вредности и дисперзија. Израчунамо:

$$t^* = \frac{\hat{\nu}_1^* - \hat{\nu}_2^*}{\sqrt{\frac{\hat{s}_1^{*2}}{n_1} + \frac{\hat{s}_2^{*2}}{n_2}}}$$

3. Поновимо корак 2 B пута, где је B велики број, како бисмо добили бутстреп узорак расподеле тест статистике под нултом хипотезом: $\{t_1^*, ..., t_B^*\}$.

4. P-вредност израчунамо као проценат индекса i, где $i \in \{1, ..., B\}$, за које важи $|\nu_i^*| > |T|$.

5. Донесемо одговарајући закључак на основу p-вредности и датог прага значајности теста.

Са малим модификацијама, сличан приступ може да се искористи за тестирање асиметричних алтернативних хипотеза.

Напоменућемо да, иако не очекујемо да мера присутности предмета једног документа има нормалну расподелу, очекујемо асимптотску нормалност њене узорачке средине. Разлог за ово је што мера присутности предмета представља тежинску суму присутности термина, које, као што смо напоменули, саме по себи представљају вероватноће. Ово повлачи ограниченост присутности предмета на интервал $[0, 1]$ и даје услове за централну граничну теорему. Класичан Студентов тест можемо да користимо у случају великих обима узорака.

Глава 6

Примена

6.1 Рецензије женске одеће

У бази података [19] се налази 23486 анонимизираних онлајн рецензија женске одеће. Саме рецензија су дате у виду три променљиве: текстуална рецензија, оцена (цео број од 1 до 5), и препорученост (бинарна вредност, индикатор да ли рецензент препоручује производ). Уз то, постоји неколико пропратних променљивих (старост рецензента, категорија производа, итд.) које нећемо даље користити.

У овом одељку ћемо да анализирамо ову базу тако што ћемо да направимо модел над њом и користимо статистику присутност предмета као меру позитивног сентимента.

Претпроцесуирање

За прављење LDA модела нам је из базе потребан само текстуални облик рецензија. Основни циљ је да се чисти текстуални подаци трансформишу у облик који је потребан моделу, тј. матрицу докумената и термина. Пре него што то урадимо, вршимо неке честе кораке у трансформацији текста (в. [12]):

1. Сва слова претварамо у мала слова: Циљ овог корака је да велика слова на почетку реченице, насумична велика слова и разлике у капитализацији у неформалном тексту не доведу до разликовања једног истог термина. Иако ово може потенцијално да доведе до изједначавања речи које нису исте, нпр. енг. Smith као презиме и енг. smith као ковач, сма-

трамо да овакви случајеви нису чести у нашем тексту и да су предности јаче од мана.

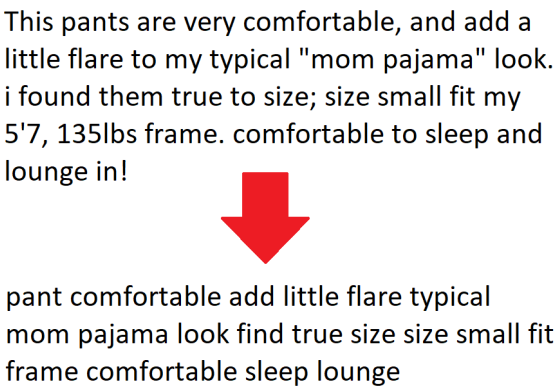
2. Бришемо бројеве из текста: Третирање речи као категоричке променљиве онемогућава смислен начин обраде нумеричких података, те уклањамо све речи које у себи садрже цифре.

3. Бришемо знакове интерпункције: Знакови интерпункције имају битну улогу у општем моделирању језика, али губе смисао када се ред речи не узима у обзир.

4. Бришемо стоп речи: Стоп речи представљају ручно направљен скуп речи одређеног језика. То су углавном честе речи које саме по себи, без контекста, не носе значајну информације. Углавном су то речце, предлози и заменице. Слично као и за знакове интерпукције, њихов језички значај се губе када се не разматра ред речи, па могу само да додају шум међу битне податке.

5. Лематизујемо речи: Лематизација представља свођење речи на њен основни облик, тј. уклањање промена по падежима, временима, множини, итд. Суштина овог корака је да спречи разликовање истог термина у различитим облицама.

6. Уклањамо ретке речи. Тематска припадност речи не може прецизно да се научи у ситуацијама када не постоји довољно докумената у којима се реч спомиње. Још једна сврха уклањања ретких речи јесте смањење димензије матрице докумената и термина, што може знатно да убрза оцену параметара. Не постоји унапред предодређена граница за уклањање, а постоје препоруке да се чак уклоне све речи које се не појављују у бар 1% свих докумената. Ми ћемо ипак узети конзервативији приступ, и уклонити само речи које се појављују у само једном документу. Једноставна природа језика у онлајн рецензијама и релативно кратка просечна дужина рецензије доводи до малог вокабулара за овај број докумената, те немамо бриге о високој димензионалности података.



Слика 6.1: Пример једне рецензије, пре и после претпроцесуирања.

LDA Модел

За оцењивање параметара модела користимо Гибсово узорковање (в. [13]). Број тема постављамо на 5, јер не очекујемо да се велики број тема разазна због специфичне природе текста. Хиперпараметре Дирихлеових расподеле постављамо на $\alpha = 0.2$, односно на $\frac{1}{K}$, као што је препоручено у [14] и $\beta = 0.05$, подразумевано вредност у имплементацији (в. [17]).

Приказујемо свих 5 тема преко 10 речи са највећом вероватноћом:

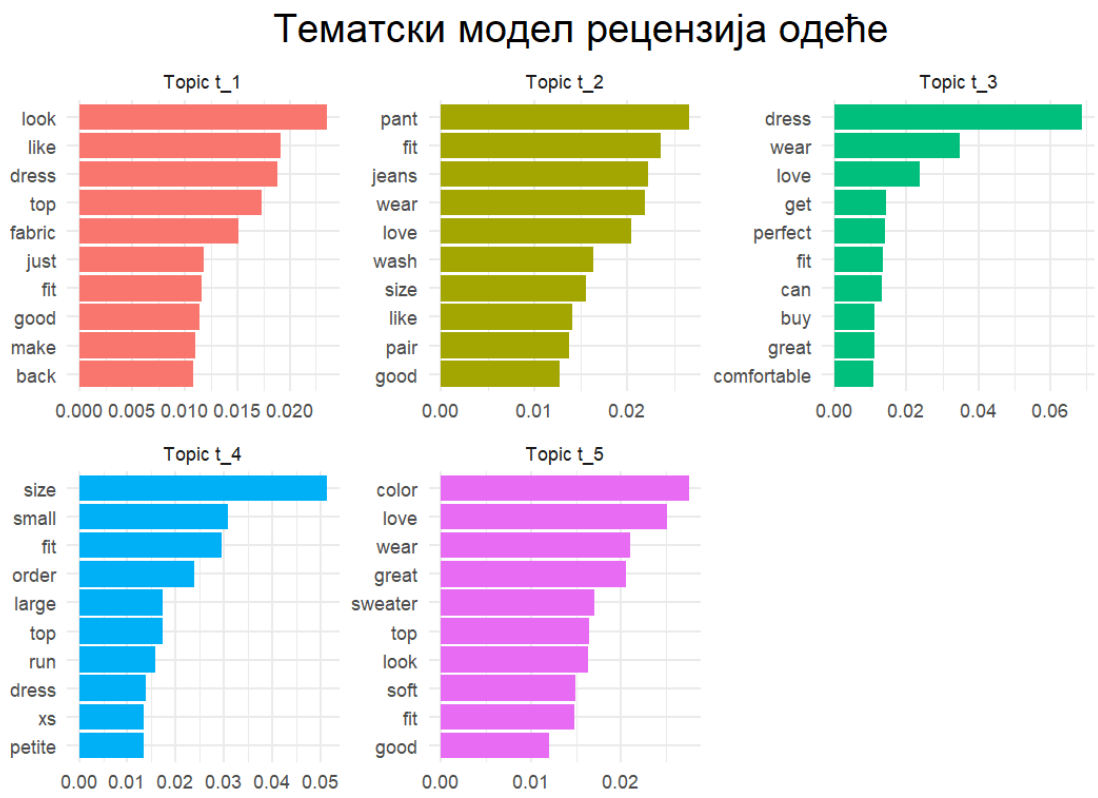
Као што видимо на слици 6.2, теме имају велика преклапање. Међу речима које се појављују у најчешћим речима више тема налазе се: dress, fit, good, love, size, итд. Ово преклапање у општем случају доводи до тежег разликовања тема, али је у нашем случају и очекивано услед уског вокабулара и тематике текстова.

Опет, разазнују се одређене теме. Јасно се види да се четврта тема бави величином и кројом. Друга тема, у односу на остале, већи утицај даје пантало-нама. Велики део речи треће теме се односи на опште позитиван сентимент. Ипак, прва и пета тема немају неке семантичке везе унутар себе. До сличног закључка долазимо гледајући кохерентност тема.

Тема	t_1	t_2	t_3	t_4	t_5
Кохерентност	0.10096	0.10456	0.08181	0.20231	0.04299

Табела 6.1: C_V кохерентност тема.

Убедљиво највећу кохерентност има четврта тема, тема о крају, а пета тема има убедљиво најмању кохерентност. Кохерентност се не слаже са нама



Слика 6.2: Теме над рецензијама женске одеће.

око прве теме, у којој не видимо јасну интерпретацију, док кохерентност није толико ниска.

Присутност позитивног сентимента

У овом одељку ћемо да искористимо присутност предмета да уочимо позитиван сентимент у рецензијама.

Термини предмета су стандрадна листа (в. [15]) кључних речи позитивног сентимента у обради природног језика. Листа је екстензивна и садржи 2006 речи. Након лематизације датих речи и брисања речи ван вокубалара, остаје 500 речи.

Присутност по темама

За почетак ћемо да израчунамо присутност позитивног сентимента у свакој од тема пронађених од стране модела. Математички, ове мере одговарају присутности документа код ког се одговарајућа тема јавља са вероватноћом 1.

Тема	t_1	t_2	t_3	t_4	t_5
Присутност	0.2847	0.3646	0.3979	0.3383	0.3611

Табела 6.2: Присутност позитивног сентимента у темама модела.

У табели 6.2 видимо да трећа тема исказује највише позитивног сентимента, што је резултат који смо очекивали.

Расподела присутности по оцени

Даље ћемо да израчунамо присутност за сваку рецензију у корпусу и приказаћемо расподелу присутности, за различите вредности оцене рецензије. Очекујемо да рецензије са вишим оценама исказује позитиван сентимент у већој мери у просеку.

Као што видимо на слици 6.3, не постоји велика разлика у расподелама присутности позитивног сентимента за три најниже оцене. Разлог за ово може бити чињеница да се суштинска разлика у негативним рецензијама не налази у количини позитивитета, већ у различитој количини негативитета. Оцене 4 и 5 се ипак јасно разликују од осталих и показују тренд померања расподеле удесно.

Тест разлике присутности за различите оцене

Сада тестирамо да ли су разлике у присутности позитивног сентимента значајне међу рецензијама различитих оцена. Тестирамо да ли постоји позитивна разлика у узаступним групама рецензијама. Конкретно, извршавамо

Слика 6.3: Хистограми густине присутности позитивног сентимента за различите вредности оцене.

четири различита теста, редом са хипотезама:

$$H_0 : \nu_2 = \nu_1 \qquad H_1 : \nu_2 > \nu_1,$$
$$H_0 : \nu_3 = \nu_2 \qquad H_1 : \nu_3 > \nu_2,$$
$$H_0 : \nu_4 = \nu_3 \qquad H_1 : \nu_4 > \nu_3,$$
$$H_0 : \nu_5 = \nu_4 \qquad H_1 : \nu_5 > \nu_4,$$

где ν_i, где $i \in \{1, 2, 3, 4, 5\}$, представља присутност позитивног сентимента у рецензијама са оценом i.

У табели 6.3 видимо да, изузев теста разлике присутности рецензија са оценом 1 и рецензија са оценом 2 (где имамо супротан закључак), постоји статистички значајна разлика у присутности позитивног сентимента између

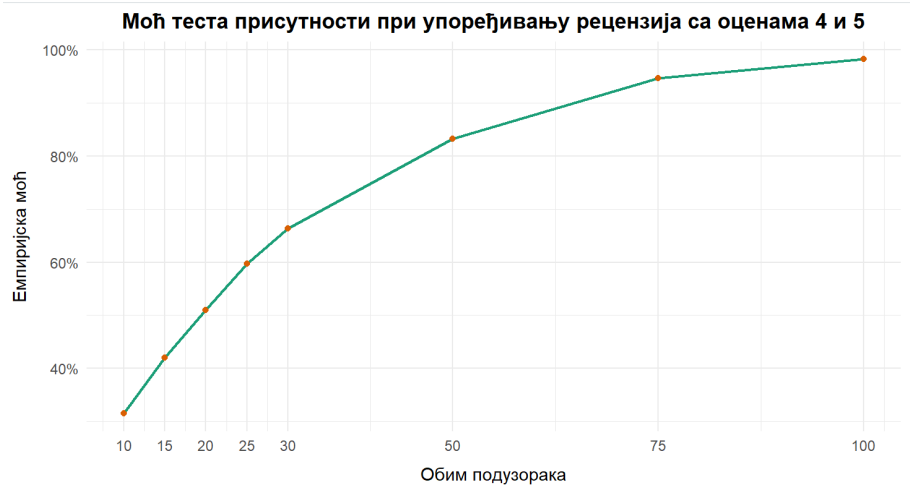
Поређење	T вредност	Прис. (више)	Прис. (ниже)	p-вредност
Оцене 2 и 1	-2.6109	0.3151	0.3110	0.9956
Оцене 3 и 2	3.4862	0.3110	0.3152	0.0003
Оцене 4 и 3	22.4374	0.3152	0.3404	0.0000
Оцене 5 и 4	31.7473	0.3404	0.3704	0.0000

Табела 6.3: Приказ резултата тестова разлике у присутности за рецензије груписане по оценама.

сваке 2 узаступне групе рецензија. Закључујемо да статистика присутности предмета може да уочи разлике у сентименту на примеру реалних података и даје валидан ненадгледан начин мерења сентимента.

Емпиријска моћ теста

За крај ћемо оценити моћ теста разлике присутности на примеру тестирања рецензија са оценом 4 и рецензија са оценом 5. Моћ оцењујемо, тако што, за различите величине узорака, извлачимо 3000 узорака рецензија са оценом 4 и оценом 5 и бележимо проценат тестова који прихватају алтернативну хипотезу $H_1 : \nu_5 > \nu_4$.



Слика 6.4: Моћ теста присутности позитивног сентимента при упоређивању рецензија са оценом 5 и рецензија са оценом 4, у зависност од величине узорка из сваке од група. Сваки тест је извршен са $B = 10000$ бутстреп понављања

Са графика на слици 6.4 видимо стрм раст моћи при малим величинама узорка, као и јасну конвергенцију моћи ка 100%, за велике узорке.

6.2 Књиге апстраката конференције CMStatistics

У овом одељку анализирамо апстракте на конференцији CMStatistics у периоду од 2008. до 2020. године. Главни циљ је да измеримо утицај развоја пројекта CRoNoS (`http://www.cronosaction.com/`) на тематику докумената. Тачније, пројекат CRoNoS, започет 2014. године, дефинише меморандум у којем су истакнути научно-истраживачки циљеви пројекта. Истражићемо да ли се од почетка пројекта значајно повећала присутност истраживачких циљева CRoNoS-а у научним радовима конференције.

LDA модел

Претпроцесуирање текста пре конструкције матрице докумената и термина смо извршили идентичним корацима као и у претходном примеру.
Одлучили смо да поставимо број тема модела на $K = 30$, како бисмо потенцијално ухватили што већи број различитих истраживачких области. Хиперпараметре α и β смо поставили редом на $1/30$, односно $\frac{1}{K}$, и 0.05, подразумевану вредност. Параметре модела смо оценили методом Гибсовог узорковања.

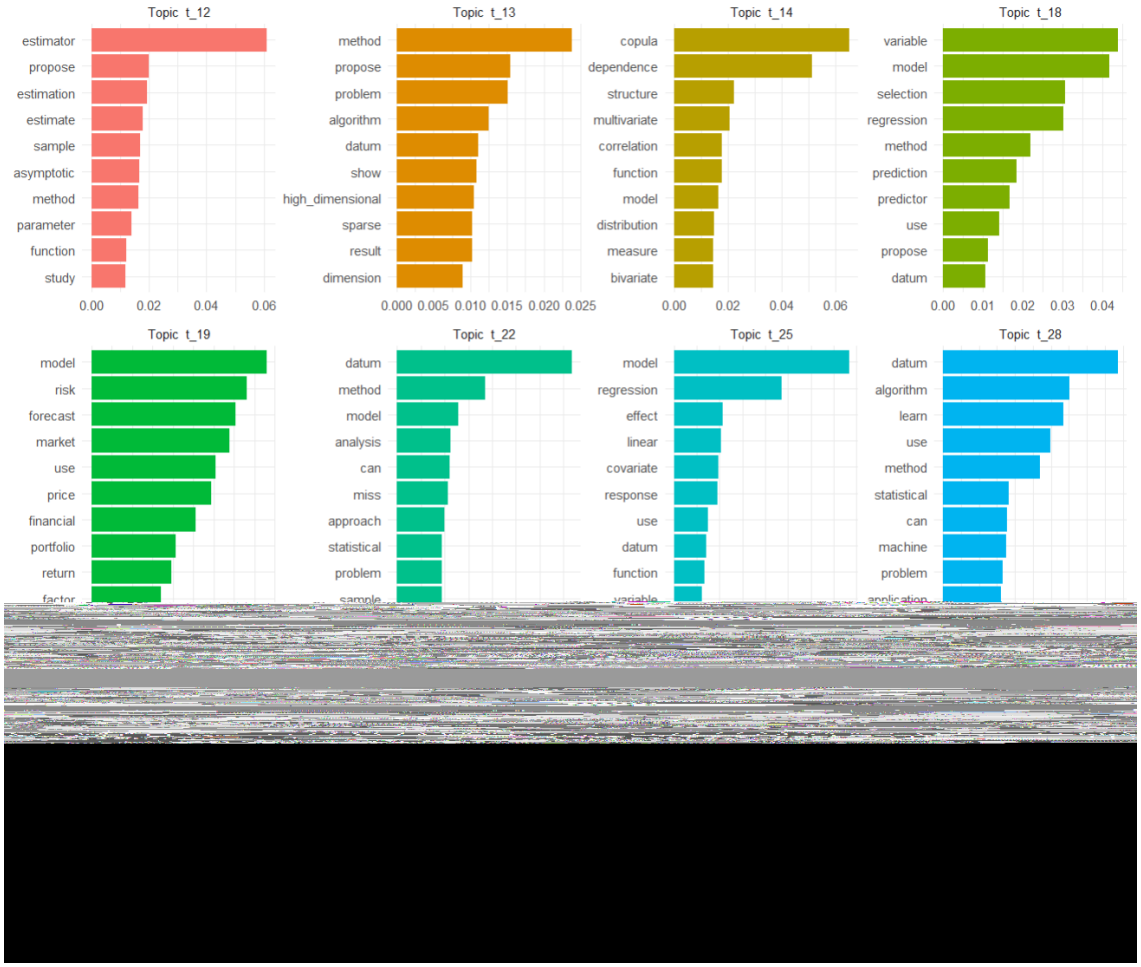
На слици 6.5 видимо узорак од 12 тема модела. Скренули бисмо пажњу на теме 12 и 19. Одокативно, тема 12 се бави оценама и успешно 3 различита облика речи estimate сврстава у ову тему. Тема 19 издваја термине финансијске математике, који се у осталим темама не појављују.

Присутност CRoNoS истраживачких тема

Како бисмо израчунали присутност, морамо дефинисати истраживачке теме CRoNoS пројекта као предмет, тј. као скуп термина са тежинама. Ово радимо у неколико корака:

1. Ручно издвојимо део меморандума CRoNoS пројекта који се тиче научних циљева пројекта.

2. Дати текст подлегнемо истим корацима предпроцесуирања које смо применили над књигама апстраката.



Слика 6.5: Приказ речи са највећом вероватноћом за 12 насумично одабраних тема модела.

3. Из добијеног текста издвојимо све јединствене речи и њихове фреквенције.

4. Фреквенције речи претворимо у релативне фреквенције, тј. скалирамо их тако да у збиру дају 1. Издвојене речи представљају термине предмета, док релативне фреквенције представљају тежине термина.

У остатку текста сваки помен присутности се односи на овако дефинисану присутност предмета. У табели 6.4 видимо присутност и C_V кохеренсност у раније приказаним темама. Убедљиво највећу кохерентност међу узоркованим темама има тема 12, која мало „вара систем", јер користи 3 различите речи истог корена. Као што смо напоменули, ово је ипак пожељно понашање.

	t_{12}	t_{13}	t_{14}	t_{18}	t_{19}	t_{22}
Присутност	1.3488	1.9893	0.8243	1.8205	1.160	2.7654
Кохерентност	0.5664	0.1479	0.2372	0.2165	0.327	0.1673

	t_{25}	t_{28}	t_3	t_5	t_8	t_9
Присутност	1.8319	2.5775	1.4936	1.5539	2.8271	1.50685
Кохерентност	0.1958	0.2865	0.1645	0.1174	0.1882	0.09527

Табела 6.4: Присутност CRoNoS-а и C_V кохерентност раније приказаних тема.

Високу кохерентност такође има и тема 19, тема о финансијама. Највећу присутност међу узоркованим темама постижу теме 8, 22 и 28. Тема 28 спомиње машинско и статистичко учење, са поменом примена, па је висока присутност и била за очекивати. Најмању пристуност има тема 14, која се већински састоји од чисто математичких појмова, а ниједна од главних речи не спада у примене.

Тест ефекта CRoNoS пројекта

Тестираћемо да ли се након почетка CRoNoS пројекта, просечна пристуност CRoNoS истраживачких тема у радовима повећала. Циљ је да се покаже да је присутност порасла под утицајем пројекта. Претпостовљамо да постоје 2 потпопулације докуманата - радови пре пројекта, односно радови објављени пре 2014. године и радови након почетка пројекта (од 2014. године надаље). Корпус делимо по овом критеријуму на два подузорка и тестирамо да ли је присутност друге групе већа, тј. тестирамо:

$$H_0 : \nu_{2014-2020} = \nu_{2008-2013}$$
$$H_1 : \nu_{2014-2020} > \nu_{2008-2013}$$

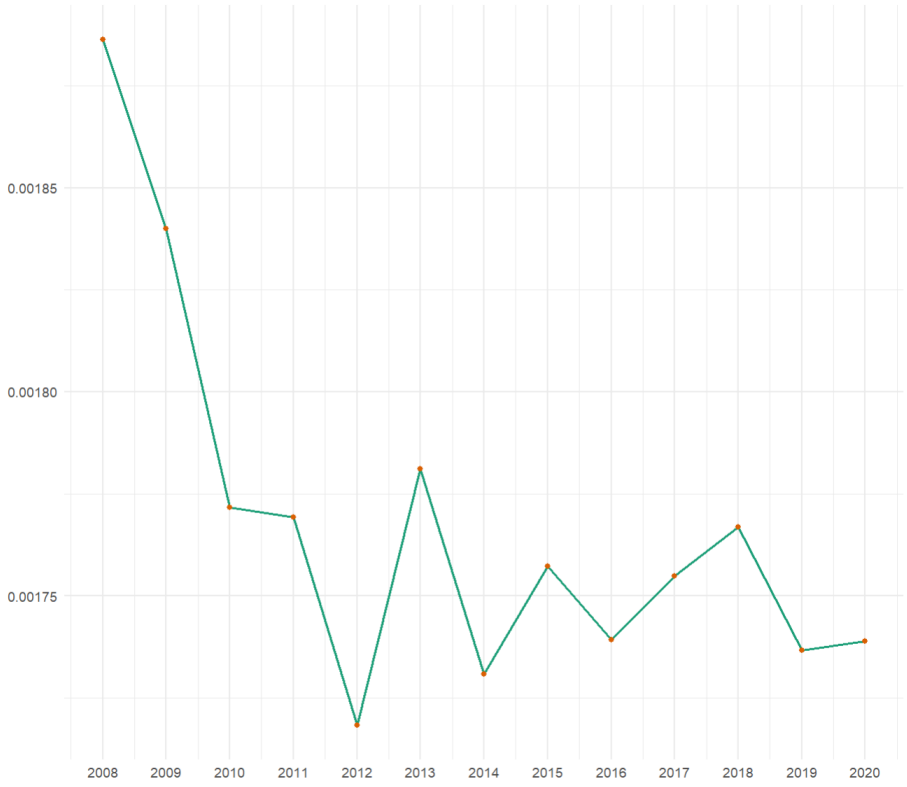
T вредност	Присутност (2014-2020)	Присутност (2008-2013)	p-вредност
-2.9043	1.7469	1.7741	0.9983

Табела 6.5: Резултати тестирања повећања присутности CRoNoS истраживачких тема у радовима на CMStatistics конференцији након почетка CRoNoS пројекта.

У табели 6.5 видимо резултате овог теста. Помало изненађујуће, не само да не постоји значајно повећање, већ нам p-вредност говори да је присутност

значајно већа у периоду пре пројекта. Напоменимо да ово не значи да је пројекат био неуспешан. Скала конференције се значајно повећала од почетка пројекта, те је могуће да је то довело до разноликијих тема, разводњавајући утицај радова са кључним темама у просеку.

Напомена 6.2.1 *Вредности присутности се не поклапају потпуно са резултатима приказаним у раду [18] због другачије методе екстракције кључних термина (и њихових тежина) CRoNoS пројекта. Закључци су ипак исти.*



Слика 6.6: Присутност CRoNoS тема кроз кроз године.

На слици 6.6 видимо да је нагли пад присутности забележен у прве три године одржавања конференције. Ово се уклапа у хипотезу да су, при раном ширењу конференције, теме у радовима постале разноликије.

Глава 7

Закључак

Пробабилистички тематски модели нам дају моћан алат за анализу текста. Проналазе интерпретабилне теме, успешно уочавајући латенту структуру текста. Мера присутности представља интерпретабилну меру која нам, уз помоћ оцењених параметара ових модела, даје софистициран, робусни начин уочавања докумената који испољавају жељену тематику.

На примеру рецензија смо видели да мера присутности успешно разликује рецензије груписане по оценама. Природно лабелирана база података као што је ова нам је омогућила да валидирамо меру присутности над неидеалним реалним подацима. Сличне базе би могле да служе за даље истраживање и оптимизацију мере присутности. Слабије оцењене рецензије (оцене од 1-3) нису имале велике разлике у присутности позитивног сентимента. За уочавање ових разлика, сматрамо да је потребно и мерити негативни сентимент. Те, у будућим радовима, предлажемо да се размотри разлика присутности позитивног и негативног сентимента као мера. Осим тога, постоји простор за боље дефинисање предмета позитивног сентимента, где смо ми узели најпростији облик предмета са једнаким тежинама.

Кроз рад, све експерименте смо вршили једном, не валидирајући резултате при различитим конфигурацијама модела и различитим насумичним ефектима при оптимизацији. У вези овога смо започели прелиминарну анализу ефеката хиперпараметара модела, као и везу између кохерентности и моћи теста присутности. Наши резултати нису показали ниједну значајну везу, што са позитивне стране може да буде индикатор робусности мере присутности, а са друге доводи у питање процес одабира конфигурације тематског модела са циљем анализе присутности. Наиме, кохерентност представља дефакто

меру квалитета ове класе модела, али оптимизација кохерентности не мора да повлачи оптимизацију мере присутности. У вези овога, сматрамо да је потребно извршити дубљу анализу, и на различитим базама података.

У другом практичном примеру смо анализирали апстракте научних радова CMStatistics конференције. Валидирали смо резултате добијене у другим радовима, са другачијим процесом. Језик са којим смо се сусрели, природно има велики број синтагми, нпр. машинско учење, Монте Карло методе, Марковљеви ланци. Методе које уочавају ове синтагме, нпр. преко мере NPMI дате у [5], могу потенцијално да унапреде корак претпоцесуирања у будућој анализи. За мерење присутности истраживачких тема предвиђених CRoNoS пројектом смо, на релативно једноставан начин, издвојили кључне термине и њихове тежине из меморандума пројекта. Ово је за последицу имало појављивање нестручних кључних речи, везаних за саму организацију пројекта, са високим тежинама у предмету присутности. Напредније метода екстракције ових кључних термине би могле потенцијално да дају веродостојније резултате.

Општост мере присутности значи да може једноставно да се прилагоди већини пробабилистичких тематских модела. Сви експерименти одрађени у раду могу да се рекреирају са другачијим, модернијим моделима. Цео метод екстракције кључних термина и тестирања присутности тако добијене теме је агностичан ка специфичној бази. Процес који је детаљисан може једноставно да се примени на радове у оквиру других пројеката и са других конференција, остављајући отворена врата како за будуће, тако и за ретроспективне анализе ефекта развоја разних пројеката.

Библиографија

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[2] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[3] David M. Blei and John D. Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 113–120, 2006.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Gerlof Bouma. Normalized (Pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, 2009.

[6] Jordan Boyd-Graber, Yuening Hu, and David Mimno. *Applications of Topic Models*, volume 11. Now Publishers Inc., 2017.

[7] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. Available at `https://arxiv.org/abs/2005.14165`.

[9] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NeurIPS)*, pages 288–296. Curran Associates, Inc., 2009.

[10] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, 1990.

[11] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035, 2018.

[12] Matthew J. Denny and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189, 2018.

[13] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(Suppl 1):5228–5235, 2004.

[14] Matthew Hoffman, David M. Blei, and Francis Bach. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23 (NeurIPS)*, pages 856–864, 2010.

[15] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, Seattle, Washington, USA, August 22–25 2004.

[16] Frederick Jelinek. Continuous speech recognition by statistical methods. In *Proceedings of the IEEE*, volume 64, pages 532–556, 1976.

[17] Tommy Jones. textminer: An r package for text mining using topic models, document similarity, and other text mining tools. `https://cran.r-project.org/package=textmineR`, 2016. R package version 3.0.

[18] Louisa Kontoghiorghes and Ana Colubi. New metrics and tests for subject prevalence in documents based on topic modeling. *International Journal of Approximate Reasoning*, 157:49–69, 2023. Available online 2 March 2023.

[19] nicapotato. Women's e-commerce clothing reviews. `https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews`, 2017. Accessed: 2023-11-9.

[20] Ingram Olkin and Herman Rubin. Multivariate beta distributions and independence properties of the wishart distribution. *The Annals of Mathematical Statistics*, 35(1):261–269, 1964.

[21] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

[22] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 399–408. ACM, 2015.

[23] José Ramón Saura. A three-stage methodological process of data text mining: A ugc business intelligence analysis. *ResearchGate*, March 2019. License: CC BY-SA 4.0.

[24] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July 1948.

[25] Herb Susmann. Dirichlet distribution. `https://observablehq.com/@herbps10/dirichlet-distribution`, 2021. Accessed: 2024-08-22.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. Available at `https://arxiv.org/abs/1706.03762`.

[27] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112, 2009.

[28] Yixin Wang and David. M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

[29] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qiang V. Le, Tomas Mikolov, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. Available at `https://arxiv.org/abs/1609.08144`.

[30] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA, 1949.

[31] Aleš Žitnik and Blaž Zupan. Evaluating topic models: A comparative study. In *Proceedings of the 14th International Multiconference Information Society*, pages 81–84. Jozef Stefan Institute, 2011.