Методи машинного навчання в текстовій аналітиці

Виконав: Курінний І. В.
Керівник: к.т.н., доц. Тимощук О. Л.

Актуальність

Повсюдна інформатизація призводить до накопичення величезних об'ємів текстових даних в науці та бізнесі. Ці дані потребують відповідного аналізу для вирішення найрізноманітніших проблем.

Об'єкт та предмет дослідження

Об'єкт досліждення — методи машинного навчання в текстовій аналітиці

Предмет дослідження — семантичний аналіз та тематичне моделювання для колекцій текстових документів

Загальні поняття

Інтелектуальний аналіз тексту (Text Mining) — напрям інтелектуального аналізу даних та штучного інтелекту, метою якого є отримання інформації з колекцій текстових документів, ґрунтуючись на застосуванні ефективних, у практичному плані, методів машинного навчання та обробки природної мови.

Прикладні задачі

- Машинний переклад
- Питально-відповідальні системи
- Інформаційний пошук
- **Семантичний аналіз**
- **Розпізнавання/визначення теми**
- ...

Етапи аналізу тексту

1. **Витяг тексту**
   обробка документів у форматі pdf, html, ...
2. **Токенізація (tokenization)**
   процес розбиття тексту на слова, фрази, символи або інші елементи, які несуть смислове навантаження
3. **Видалення стоп-слів**
   це службові чатини мови, прийменники, сполучники, артиклі: a, of, on, ...
4. **Лемматизація (stemming)**
   процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс
5. **Машинне навчання**
   набір документів представляється у вигляді матриці; перетворення: tf-Idf, SVD розклад; типові методи: kNN, SVM.

Тематичне моделювання

- **Тема** — умовний розподіл на множині термінів, $p(w|t)$ — ймовірність (частота) терміна w в темі t
- **Тематика документа** — умовний розподіл, $p(t|d)$ — ймовірність (частота) теми t в документі d
- **Тематична модель** автоматично виявляє латентні теми по частотам термінів в документах $p(w|d)$

Тематичне моделювання

- **Знайти**: параметри ймовірнісної тематичної моделі

$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

- $\phi_{wt} = p(w|t)$ — ймовірність термінів w в кожній темі t
- $\theta_{td} = p(t|d)$ — ймовірність теми t в кожному документі d
- шуканий розклад — $F_{W \times D} = \Phi_{W \times T} \times \Theta_{T \times D}$

- **Принцип максимума регуляризованої правдоподібності**: оскільки рішення не є єдиним, то вводиться критерій регуляризації $R(\Phi, \Theta)$

$$\begin{cases} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \\ \sum_{w \in W} \phi_{wt} = 1 \qquad \phi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1 \qquad \theta_{td} \geq 0 \end{cases}$$

- Рішення можна знайти за допомогою EM-алгоритму, ітерації для системи рівнянь відносно змінних $\phi_{wt}, \theta_{td}, p_{twd} = p(t|d,w)$

Адитивна регуляризаці тематичних моделей

В модель можна вводити наступні регуляризатори:

- розрідження матриці Φ
- розрідження матриці Θ
- декореляція тем в матриці Φ
- згладжування матриці Φ (для фонової теми)
- балансування класів

Критерії якості

Для аналізу побудованої тематичної моделі використовуються наступні критерії:

1. **когерентність топ-слів** — міра узгодженості слів в темі
2. **перплексія** — чисельне значення перплексії не має інтерпретації і дозволяє лише порівнювати алгоритми між собою.
3. **розрідженість матриць** Φ, Θ — доля нульових елементів
4. **характеристики ядра теми**
   - чистота — сумарна ймовірність слів ядра
   - контрастність — середня ймовірність зустріти слова ядра в конкретній темі

Модельні задачі

1. Семантичний аналіз твітів та знаходження геолокації користувачів
2. Визначення теми листів з подальшим семантичним аналізом

Аналіз твітів

- **Постановка задачі**
  провести семантичний аналіз твітів та визначити в якому штаті проживає користувач
- **Практичне застосування**
  даний підхід до аналізу твітів може бути використаний при розробці маркетингових компаній, визначенні громадської думки і т. ін.

Теми листів

- **Мета**
  побудова платформи, яка б повністю замінила опитування в компаніях
- **Постановка задачі**
  провести аналіз переписки між співробітниками та побудувати тематичну модель
- **Практичне застосування**
  дозволить визначати настрої співробітників та клієнтів, найбільш обговорювані теми, знаходити проблемні місця та недоліки в організації роботи

Дані

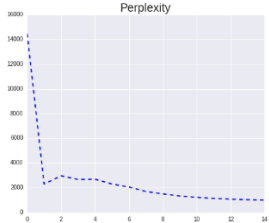

```
{
    "_id": {
        "$oid": "573c8abaa7a141000335de78"
    },
    "from_email": "rerickson@pkgcompliance.com",
    "to_email": "rerickson@pkgcompliance.com",
    "id": "154bfeef578ab22c",
    "gmail_label": "",
    "content": "Hi Carol, That sounds great! We don\u2019t have a trip
    on the books but we are definitely planning to make some more sales
    trips in June / July timeframe ...",
    "timestamp": {
        "$date": "2016-05-17T18:16:29.000Z"
    },
    "subject": "RE: 20 day aging sample"
}
```

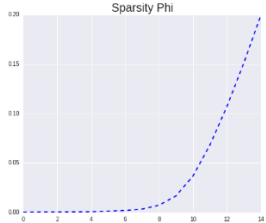
Побудова моделі

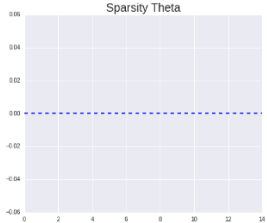
Стратегія додавання регуляризаторів

1. Перший етап: 14 проходів по колекції
   - декореляція матриці Φ
   - згладжування матриці Φ по фоновій темі
2. Другий етап: 4 проходів по колекції
   - розрідження матриці Φ; модальність content; всі теми, окрім фонової
   - розрідження матриці Φ; модальність author, time, emotion; всі теми, окрім фонової
   - розрідження матриці Θ; всі теми, окрім фонової

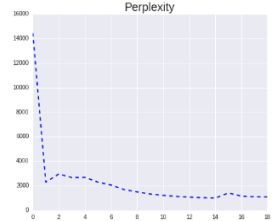
Перший етап

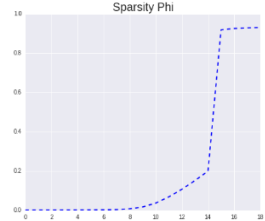


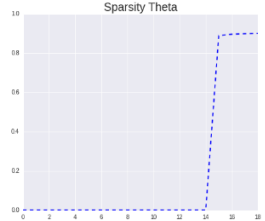




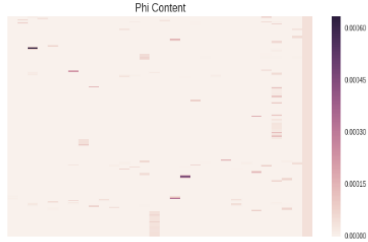
Другий етап



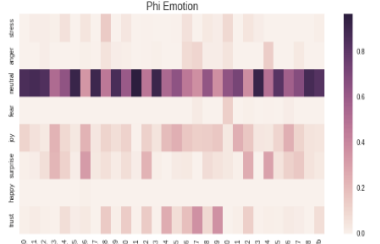




Перший етап

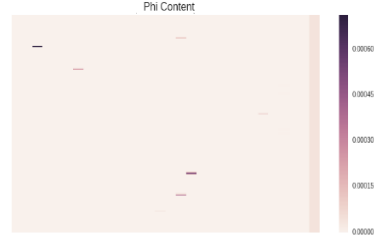

Phi Content
0.00060
0.00045
0.00030
0.00015
0.00000

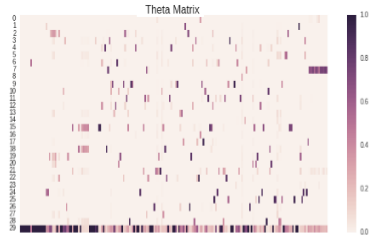

Theta Matrix
0.75
0.60
0.45
0.30
0.15
0.00

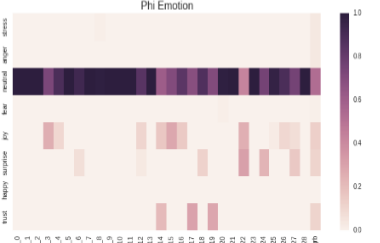

Phi Emotion
stress
anger
neutral
fear
joy
surprise
happy
trust
0.8
0.6
0.4
0.2
0.0


Phi Author
0.8
0.6
0.4
0.2
0.0

Другий етап


Phi Content
0.00060
0.00045
0.00030
0.00015
0.00000


Theta Matrix
1.0
0.8
0.6
0.4
0.2
0.0


Phi Emotion
stress
anger
neutral
fear
joy
surprise
happy
trust
1.0
0.8
0.6
0.4
0.2
0.0


Phi Author
1.0
0.8
0.6
0.4
0.2
0.0

Результати

Критерій	Перший прохід	Другий прохід
Sparsity Phi	0.199	0.931
Sparsity Theta	0.000	0.901
Kernel contrast	0.389	0.374
Kernel purity	0.821	0.793
Perplexity	989.961	1083.039
Coherence	21.035	27.402

Результати

- Матриця Фі

	sbj_0	sbj_1	sbj_2	sbj_3	sbj_4	sbj_5	sbj_6	sbj_7	sbj_8	sbj_9
partially	0	0	0	0.00036	0	0	0	0	0	0
certificate	0	0	0	0	0	0	0	0	0	0
lbs	0	0	0	0	0	0	0.00002	0	0	0
acquire	0	0	0.0016	0	0	0	0	0	0	0
platforms	0	0	0	0	0	0	0	0	0	0
russians	0	0	0	0.0001	0	0	0	0	0	0
actress	0	0	0	0	0	0	0	0	0	0
gate	0	0.0005	0	0	0	0	0	0	0	0
million	0	0	0	0.0043	0	0	0	0	0	0

	sbj_{10}	sbj_{11}	sbj_{12}	sbj_{13}	sbj_{14}	sbj_{15}	sbj_{16}	sbj_{17}	sbj_{18}	sbj_{19}
partially	0	0	0	0	0	0	0	0	0	0
certificate	0	0	0.001	0	0	0	0	0	0	0
lbs	0	0	0	0	0	0	0	0	0	0
acquire	0	0	0	0	0	0	0	0	0	0
platforms	0	0	0	0	0	0	0	0	0	0.0016
russians	0	0	0	0	0	0	0	0	0	0
actress	0	0	0	0	0	0	0	0	0	0
gate	0	0	0	0	0	0	0	0	0	0
million	0	0	0	0	0	0	0	0	0	0

Резултати

- Матрица Тета

	d_1	d_2	d_3	d_4	d_5	d_6
sbj_0	0	0	0	0.005	0	0
sbj_1	0	0	0	0.003	0	0.053
sbj_2	0.032	0.504	0	0.024	0	0.018
sbj_3	0	0	0	0	0	0.018
sbj_4	0.021	0.153	0	0.013	0.082	0
sbj_5	0	0	0	0.005	0	0.036
sbj_6	0.755	0	0.011	0.042	0	0.098
sbj_7	0.008	0	0	0.005	0.786	0.036
sbj_8	0.011	0.245	0	0.01	0.111	0.249
sbj_9	0	0	0.989	0.026	0	0
sbj_{10}	0	0	0	0	0	0.213
sbj_{11}	0.001	0	0	0.068	0.003	0.018
sbj_{12}	0.011	0	0	0.685	0	0.08
sbj_{13}	0.086	0.098	0	0	0	0
sbj_{14}	0.024	0	0	0.018	0	0
sbj_{15}	0	0	0	0.053	0	0.036
sbj_{16}	0.042	0	0	0.005	0	0.027
sbj_{17}	0.008	0	0	0.005	0	0.049
sbj_{18}	0	0	0	0.016	0	0.009
sbj_{19}	0	0	0	0.015	0.019	0.062

Результати

```
sbj0:  domain app assets api templates landing css user changes test file png accounts directly added
sbj1:  view job day ago software engineer salaries jobs change days development manager mobile month market
sbj2:  com account www follow company andersen details settings facebook area log pay notification updates program
sbj3:  read linkedin people votes network profile trademarks question corporation registered does answer mountain world logo
sbj4:  product year check support search university design website reviews social place results happy events schedule
sbj5:  grand rapids analyst business learn apply good systems easily posted match lead senior application meijer
sbj6:  message information use subject mail intended thanks recipient forward confidential sender attachments error reply direct
sbj7:  think money feel person connect point talk quick document thought inform custom seller est idea
sbj8:  like date request link following friday client site task server process come description got feedback
sbj9:  div class mailto font span style target href size color blank gmail value left family
sbj10:  data michigan management suite meetup phone property real interested space investment big customers report text
sbj11:  service privacy receive online credit offer available services code center sign terms microsoft security office
sbj12:  moneysign contact project store payment card order price number items purchase valid total save gift
sbj13:  group emails send visit marketing wrote receiving post board stop options amawm event west subscribed
sbj14:  new review just click comment customer yotpo public write great set using form written respond
sbj15:  unsubscribe address policy street home join book san having life preferences tech live register food
sbj16:  google add update access able used usa password browser info issue ticket problem coming recent
sbj17:  week solutions research content survey personal regards possible students agreement inclusion later non version paper
sbj18:  experience looking products based future learning years provide krier amazon prior minutes global required energy
sbj19:  time need work make want way start right going working things better plan really meet
```

Результати

hey anthony, good to see your email! glad to see you're working on something exciting. thanks for the details on your product. i like the name sixsents... sort of like sixth sense. i'll check it out during this weekend and let you know! more later, phani sent by outlook for android from: anthony michael bird sent: friday, august 14, 14:43 subject: quick question to: phani mantravadi hey phani, long time no talk! hope all is well. i am launching a new tool (www.sixsents.com) and wanted to get some of your feedback on it. see if you'd be interested or if you knew anyone or contacts at organizations who would be interesting in using it. or even if you simply had any critical feedback on it. basically, it hooks into any form of communication you'd like (email, chat, project management, crms, internal tools etc - besides texts, that'll be a later feature) and gives you a real time read out of the emotions of whoever is in correspondence. so if your client is stressed out, it's probably not a good time to pitch him. or how your employees are feeling about your company at any given moment and over time. i'd really appreciate your thoughts, trying to feedback from friends and colleagues.
if you'd be kind enough to forward it around, that would be fantastic as well.
thanks! tony bird

0	0	0.037	0.037	0.028	0	0.037	0.271	0.131	0.009
sbj_0	sbj_1	sbj_2	sbj_3	sbj_4	sbj_5	sbj_6	sbj_7	sbj_8	sbj_9

0.037	0.009	0.019	0.028	0.028	0	0	0.093	0.028	0.206
sbj_{10}	sbj_{11}	sbj_{12}	sbj_{13}	sbj_{14}	sbj_{15}	sbj_{16}	sbj_{17}	sbj_{18}	sbj_{19}

Висновки

В даній роботі були описані та досліджені найважливіші ідеї та принципи, які лежать в основі алгоритмів, що використовуються для розв'язання задач текстової аналітики, зокрема семантичний аналіз та тематичне моделювання.

У подальшому можливе використання досліджених методів для розв'язання різноманітних задач, які виникають при обробці та аналізі колекцій текстових документів.

Були опубліковані тезиси на конференції CAIT.

Дякую за увагу!