Word2vec и Doc2vec, кластеризация и поиск гиперпараметров

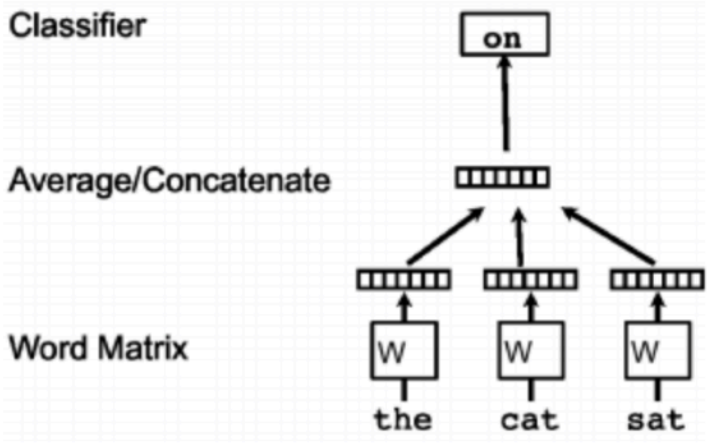
Работу выполнил Чужмаров Денис

Описание алгоритмов

Word2vec

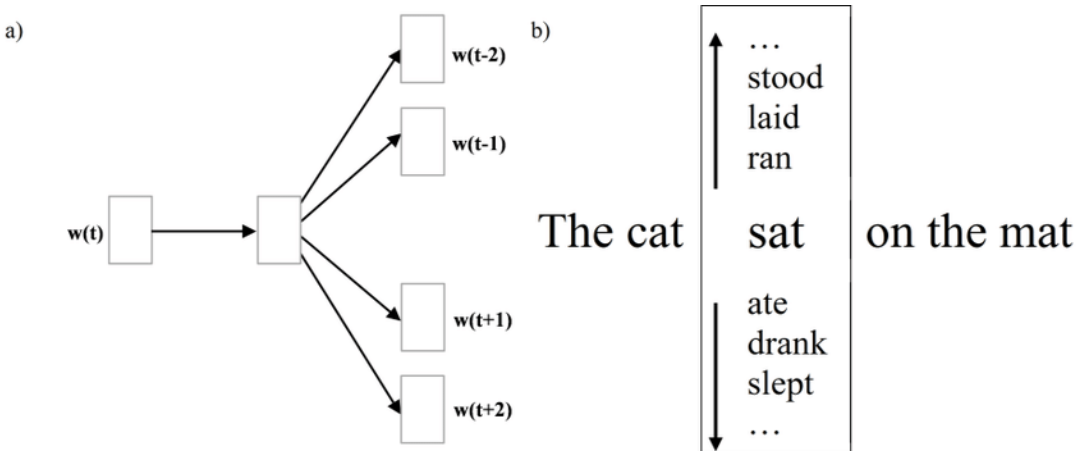
Word2vec - числовое представление для каждого слова, которое сможет отразить такие отношения, как смысловая схожесть слов.
Word2vec создается с использованием 2 алгоритмов: Continuous Bag-of-Words model (CBOW) и the Skip-Gram model.

CBOW: создает скользящее окно вокруг текущего слова, чтобы предсказать его по «контексту» — окружающим словам. Каждое слово представлено в виде вектора признаков. После обучения эти векторы становятся векторами слов.



Как было сказано ранее, векторы, представляющие похожие слова, близки по разным метрикам расстояния и дополнительно содержат числовые отношения, такие как: король - королева = человек.

Skip-Gram: в отличие от CBOW принимает слово и предсказывает окружающие слова.



Обучение:
Чтобы обучить нейронную сеть, нужно подать пары слов, найденные в наших учебных документах. Пары слов мы получаем из скользящего окна по обучающим тестам.
Сеть будет изучать статистику по количеству появлений каждой пары. Так, например, сеть, вероятно, получит гораздо больше обучающих образцов («Советский», «Союз»), чем («Советский», «Снег»). Когда обучение закончится, если вы дадите ему слово “Советский” в качестве входных данных, то он выведет гораздо более высокую вероятность для “Союза” или “России”, чем для “Снега”.

Doc2vec

Цель doc2vec - создать числовое представление документа, независимо от его длины. Но в отличие от слов, документы не имеют логических структур, таких как слова, поэтому необходимо найти другой метод.
Концепция, которую использовали Mikilov и Le, была простой, но умной: они использовали модель word2vec и добавили еще один вектор (идентификатор документа). Такая модель называется Distributed Memory version of Paragraph Vector (PV-DM).

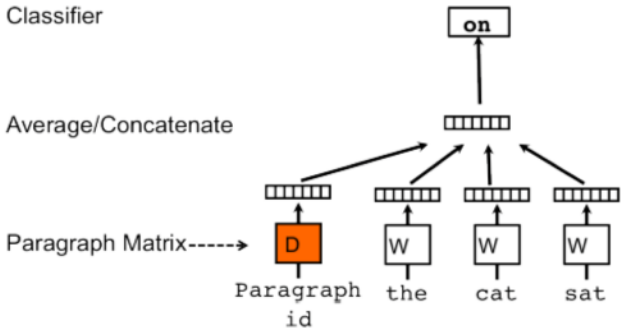


fig 3: PV-DM model

PV-DM это небольшое расширение для модели CBOW. Но вместо того, чтобы использовать только слова для предсказания следующего слова, мы также добавили еще один вектор функций, уникальный для документа.
Таким образом, при обучении векторов слов W также обучается вектор документа D, и в конце обучения он содержит числовое представление документа. В то время как векторы слов представляют концепцию слова, вектор документа предназначен для представления концепции документа.

PV-DBOW: Как и в word2vec, в другом алгоритме, аналогичном skip-gram, может быть использована версия Distributed Bag of Words version of Paragraph Vector (PV-DBOW).

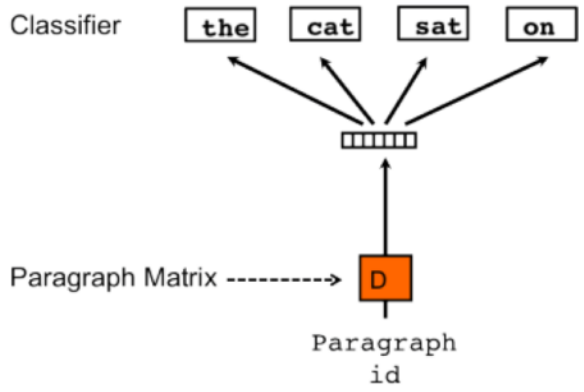


fig 4: PV-DBOW model

Такой алгоритм работает быстрее чем word2vec и использует меньше памяти, так как нет необходимости сохранять векторы слов.

Модели doc2vec могут использоваться следующим образом: для обучения требуется комплект документов. Для каждого слова создается вектор слов W, а для каждого документа-вектор документов D. Модель также тренирует веса для скрытого слоя softmax. Этот алгоритм хорошо написан в библиотеке gensim, которую мы использовали в этом исследовании.

Гиперпараметры:
- vector size - размер вектора слова, который кодирует слова (и сам документ)
- window size - размер окна, скользящего по тексту.

Кластеризация документов

Чтобы сгруппировать документы, мы будем использовать алгоритм Mini-batches K-means algorithm. Этот вариант K-means использует случайные выборки входных данных, чтобы сократить время, необходимое для обучения. Положительным моментом является то, что он использует ту же целевую функцию, что и исходный алгоритм, поэтому на практике результаты немного хуже, чем у K-средних.

Чтобы проверить качество кластеризации мы будем использовать Silhouette Coefficient. Этот коэффициент является оценочным показателем, часто используемым в задачах, где метки истинности неизвестны. Он рассчитывается с использованием среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера и изменяется от -1 до 1. Четко определенные кластеры приводят к положительным значениям этого коэффициента, в то время как неправильные кластеры приведут к отрицательным значениям. Если вы хотите узнать об этом больше, ознакомьтесь с документацией scikit-learn.

10 лучших кластеров для word2vec:

```
For n_clusters = 50
Silhouette coefficient: 0.11
Inertia:3567.650253048523
Silhouette values:
    Cluster 25: Size:49 | Avg:0.40 | Min:0.06 | Max: 0.59
    Cluster 2: Size:139 | Avg:0.31 | Min:-0.04 | Max: 0.50
    Cluster 17: Size:90 | Avg:0.27 | Min:-0.10 | Max: 0.50
    Cluster 19: Size:96 | Avg:0.27 | Min:-0.01 | Max: 0.45
    Cluster 28: Size:105 | Avg:0.25 | Min:-0.05 | Max: 0.48
    Cluster 30: Size:57 | Avg:0.24 | Min:-0.06 | Max: 0.48
    Cluster 48: Size:78 | Avg:0.23 | Min:-0.07 | Max: 0.43
    Cluster 32: Size:59 | Avg:0.22 | Min:-0.00 | Max: 0.44
    Cluster 23: Size:80 | Avg:0.20 | Min:0.03 | Max: 0.40
    Cluster 13: Size:123 | Avg:0.19 | Min:-0.06 | Max: 0.40
```

10 лучших кластеров для doc2vec:

```
For n_clusters = 50
Silhouette coefficient: 0.02
Inertia:8569.727992439684
Silhouette values:
    Cluster 33: Size:7 | Avg:0.14 | Min:0.02 | Max: 0.29
    Cluster 19: Size:387 | Avg:0.12 | Min:0.03 | Max: 0.26
    Cluster 37: Size:149 | Avg:0.10 | Min:-0.06 | Max: 0.24
    Cluster 38: Size:127 | Avg:0.09 | Min:-0.16 | Max: 0.29
    Cluster 6: Size:375 | Avg:0.08 | Min:0.00 | Max: 0.19
    Cluster 31: Size:104 | Avg:0.07 | Min:-0.15 | Max: 0.28
    Cluster 43: Size:46 | Avg:0.07 | Min:-0.02 | Max: 0.19
    Cluster 9: Size:207 | Avg:0.07 | Min:-0.08 | Max: 0.22
    Cluster 24: Size:183 | Avg:0.06 | Min:-0.11 | Max: 0.21
    Cluster 27: Size:275 | Avg:0.06 | Min:-0.04 | Max: 0.21
```

Ниже представлен пример word2vec кластера, в который алгоритм отнес статьи где говорится о преступлениях людей:

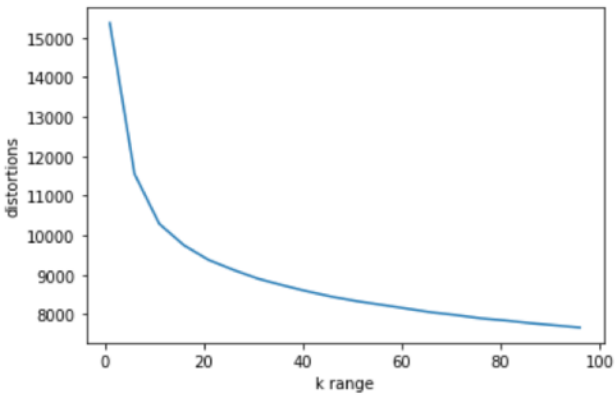
```
Fox News Poll Shows Trump Losing to Every Democratic Frontrunner - Newsweek | Fox News Poll Shows Trump Losing to E
very Democratic Frontrunner Newsweek The polls are in: here's who won — and lost — last week's debate Vox.com One b
ig lesson from the General Motors strike NJ.com Electability becomes a self-fulfilling prophecy The Washing… | The
latest Fox News poll about the 2020 election shows President Donald Trump losing to every Democratic frontrunner in
cluding Joe Biden, Bernie Sanders and Elizabeth Warren.
The survey, which was conducted from September 15 to September 17, found that Biden… [+2580 chars]
--------------
Presidential resumes: We asked voters the qualities they want most in a president, and these are the 2020 Democrati
c candidates who look best on paper | Earlier this summer, Insider polled voters on what accomplishments made them
more likely to vote for a presidential candidate. Democratic voters' most favored qualities included having release
d tax returns, gubernatorial experience, Congressional experience, … | More than half of all US presidents have bee
n lawyers. Many were born into wealth. But according to voters polled by Insider, those aren't terribly desirable t
raits for those running for president.
 Voters said they best liked candidates who had served as g… [+2278 chars]
--------------
Expectations of Andrew Yang have doubled since the last debate, according to a new poll | A new Insider poll found
Democratic voters' confidence in Andrew Yang 's debate performance has doubled since the last debate. By proportion
, that's the most significant boost out of all the 2020 contenders between July's pre-debate polling and now. Yang
and … | Entrepreneur Andrew Yang may not be a frontrunner in the 2020 presidential race, or even close, but America
ns' expectations of him are rising fast.
 According to a new Insider poll, confidence in Yang's debate performance has doubled among both self-reporte… [+36
27 chars]
```

Ниже представлен пример doc2vec кластера, в который алгоритм отнес статьи где говорится о политике:

```
Police: No evidence of shooting at northern Virginia mall | Get breaking national and world news, broadcast video c
overage, and exclusive interviews. Find the top news online at ABC news. | Authorities in northern Virginia say the
y have found no evidence that a shooting occurred at a popular mall.
The Arlington County Police Department tweeted Saturday night that authorities were continuing to conduct a search
at the Ballston Quarter mall in Ar… [+162 chars]
--------------
US woman arrested at Manila airport with baby hidden in bag | Get breaking national and world news, broadcast video
coverage, and exclusive interviews. Find the top news online at ABC news. | An American woman who attempted to carr
y a 6-day-old baby out of the Philippines hidden inside a sling bag has been arrested at Manila's airport and charg
ed with human trafficking, officials said Thursday.
They said Jennifer Talbot was able to pass through t… [+1496 chars]
--------------
Jury selection begins in killing of pregnant woman, officer | Get breaking national and world news, broadcast video
coverage, and exclusive interviews. Find the top news online at ABC news. | Jury selection has begun for a man accu
sed of killing his pregnant ex-girlfriend and a Florida police officer.
The Orlando Sentinel reports that Circuit Judge Leticia Marques told potential jurors Friday that if selected for t
he 12-member panel, they would b… [+618 chars]
```

Поиск гиперпараметров

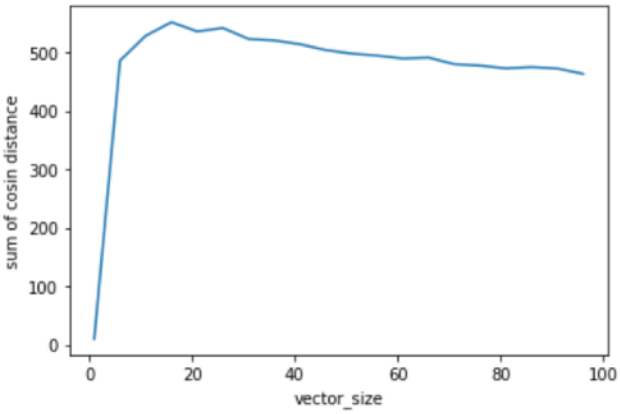
Для начала подберем гиперпараметры для алгоритма кластеризации. Поскольку мы используем kmeans, нам нужно подобрать такое k при котором кластеры отделяются лучше всего.



k в KMeans определяется методом локтя: выбираем к при котором ошибка перестает быстро уменьшаться. Из графика видно, что от 1 до 30 кластеров ошибка уменьшается давольно быстро, поэтому рекомендуемое значение к = 30.

Теперь перейдем к подбору гиперпараметров для алгоритмов word2vec и doc2vec. Как было сказано ранее, гиперпараметров у этих алгоритмов 2:
- vector size - размер вектора слова, который кодирует слова (и сам документ)
- window size - размер окна

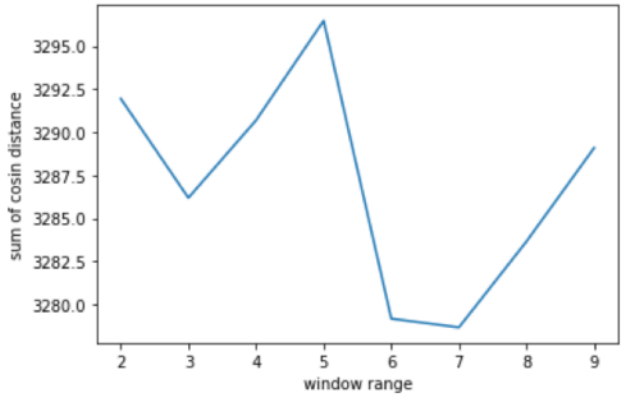
Для проверки качества алгоритма мы решили сравнивать косинусное расстояние между вектором темы и вектором текста документа. Для этого мы заново разбили базу данных и запустили поиск кипер параметров по сетке:



Из графика видно что ошибка растет от 1 до 15. После 30 начинает постепенно снижаться. И такой тренд продолжается до бесконечности. Поэтому оптимальным значением будет вектор размеров от 30-100.
Теперь посмотрим на зависимость от размера окна:

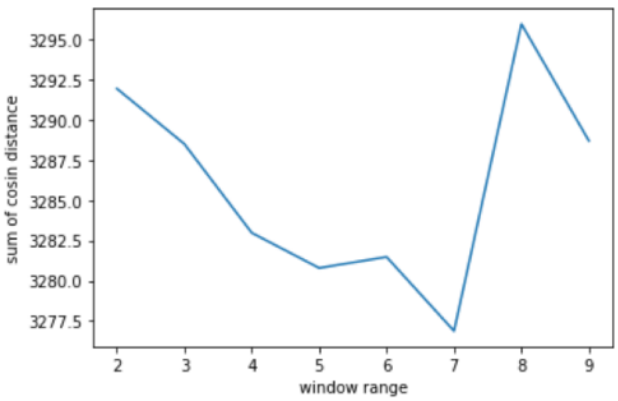
```
vector size = 30

Text(0, 0.5, 'sum of cosin distance')
```




```
vector size = 40

Text(0, 0.5, 'sum of cosin distance')
```



Левый график при размере вектора = 30, правый размер вектора = 40. В обоих случаях минимальная ошибка достигается при window size = 7. Поэтому рекомендуемое значение 7.

Другие алгоритмы

1. Методы BOW
   1.1. простой BOW
   1.2. BOW с леммами слов
   1.3. BOW с леммами и очисткой стопслов
   1.4. LDA
2. Методы, использующие эмбединги токенов
   2.1 Среднее по эмбедингу всех слов
   2.2 Среднее по эмбедингу с очисткой стоп слов
   2.3 Среднее по эмбедингу с весами tf-idf
3. Languade Models
   3.1 Language Model on embedings
   3.2 Language Model on index
4. BERT
   4.1 rubert_cased_L-12_H-768_A-12_pt
   4.2 ru_conversational_cased_L-12_H-768_A-12_pt
   4.3 sentence_ru_cased_L-12_H-768_A-12_pt
   4.4 elmo_ru-news_wmt11-16_1.5M_steps
   4.5 elmo_ru-wiki_600k_steps
   4.6 elmo_ru-twitter_2013-01_2018-04_600k_steps
5. Автоэнкодеры
   5.1 Автоэнкодер embedings -> embedings
   5.2 Автоэнкодер embedings -> indexes
   5.3 Автоэнкодер архитектура LSTM -> LSTM
   5.4 Автоэнкодер архитектура LSTM -> LSTM -> indexes
6. Эмбединги на Transfer Learning
   6.1 Эмбединги на BOW
   6.2 Эмбединг на LSTM + MaxPooling
   6.3 Эмбединг на LSTM + Conv1D + AveragePooling
   6.4 Эмбединг на LSTM + Inception + Attention
7. Triplet loss
   7.1 Triplet loss на BOW
   7.2 Triplet loss на embedings

https://habr.com/ru/post/515036/

Ссылки:

Мой gitHub: https://github.com/Res0nanceD/Summer-practice

Источники информации:

https://dylancastillo.co/nlp-snippets-cluster-documents-using-word2vec/
https://habr.com/ru/post/515036/
https://arxiv.org/abs/1405.4053
https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
https://habr.com/ru/post/446530/