Векторные представления слов

Юлия Пономарева
Data Scientist

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

Цели занятия

1. Изучим методы предобработки текста с помощью pymorphy и nltk
2. Обучим модели CountVectorizer и TfIdfVectorizer
3. Поработаем с моделью Word2Vec
4. Решим задачу классификации текста с помощью предобученной модели FastText

План занятия

1. Лемматизация в PyMystem и PyMorpy
2. CountVectorizer и TfIdfVectorizer
3. Word2Vec в gensim
4. Предобученные модели
5. Обучение нейросети на классификацию текста
6. Итоги занятия

Предобработка текстов

Терминология

МФТИ

- **Токен** - единица, на которую разбиваем текст (символ, слово, предложение и т.д.)
- **Словарь** - набор уникальных токенов
- **Документ** - набор токенов, которые принадлежат одной смысловой единицы (предложение, комментарий и т.д.)
- **Корпус** - набор всех документов

Предобработка текстов

МФТИ

- **Нормализация**: процесс приведения текста к единому формату, например, приведение всех символов к нижнему регистру.

- В некоторых задачах (**именованные сущности**) регистр важен
- В некоторых – лучше привести к одному регистру

Предобработка текстов

МФТИ

- **Удаление шума**: процесс удаления нежелательных символов, таких как знаки препинания, цифры, специальные символы и т.д.

- Слова разделены пробелами / группами пробелов / табуляциями / концом строки / пробельными символами / знаками пунктуации
- Часто приходится использовать регулярные выражения
- Готовые списки знаков пунктуации для разных языков
- Типичная путаница дефисов, минусов, коротких и длинных тире
- Для решения некоторых задач пунктуация мешает, для решения других задач (**анализ тональности**) пунктуация помогает
- Смайлики не в виде эмодзи :-) :-(

Предобработка текстов

МФТИ

- **Удаление стоп-слов** (слова, которые не несут смысловой нагрузки, например, "и", "в", "на", "как", "это" и т.д.).

- Удаление стоп-слов: союзы, предлоги, местоимения и др.
- Готовые списки стоп-слов для разных языков
- Во многих задачах стоп-слова лучше не отбрасывать, особенно в словосочетаниях

Предобработка текстов

МФТИ

- **Лемматизация**: процесс приведения слова к нормальной форме (лемме). Например, слова "бежать", "бегу", "бежали" будут приведены к лемме "бежать". Типично для русского языка.
- **Стемминг**: процесс приведения слова к его псевдооснове (отсечением окончаний). Например, слова "бежать", "бегу", "бежали" будут приведены к основе "беж". Типично для английского языка.

Мешок слов

OneHotEncoder

1. Слово -> Число.
   мама - 1, мыть - 2, рама - 3, …
   Мама мыла раму. -> [1, 2, 3]

2. Слово -> бинарный вектор
3. Слово -> частотный вектор

	are	call	from	hello	home	how	me	money	now	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	0	1	0	1	0	0	1	0	0	2	0
2	0	1	0	0	0	0	1	0	1	0	0	0
3	0	1	0	1	0	0	0	0	0	1	0	1

TfIdf

TF (term frequency — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

IDF (inverse document frequency — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF.

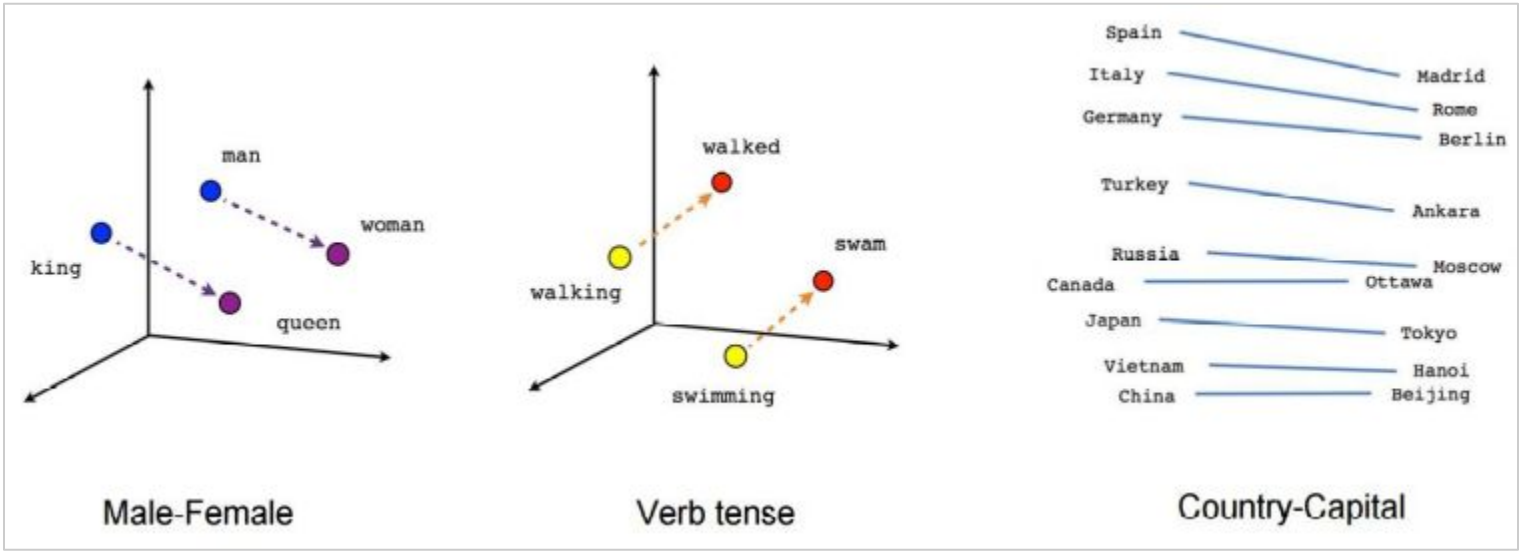
$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}$$

$$\mathbf{idf}(t, D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

МФТИ

Практика
(предобработка и BoW)

Векторная математика



vector[Queen] = vector[King] - vector[Man] + vector[Woman]

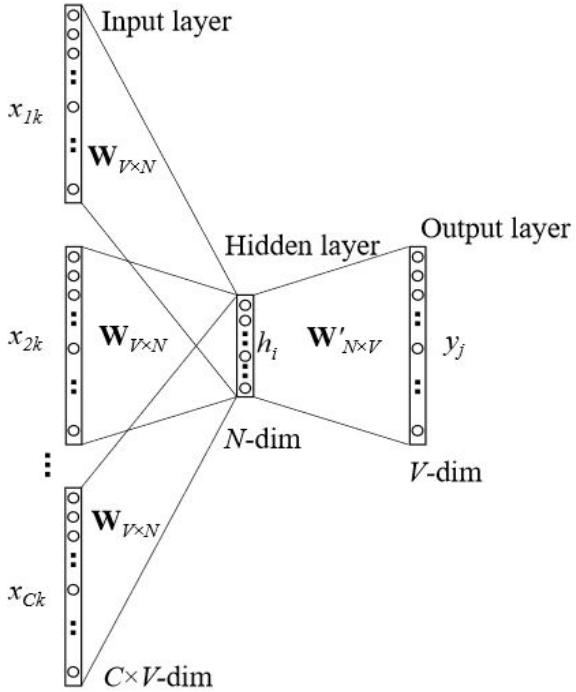
Сравнение

МФТИ

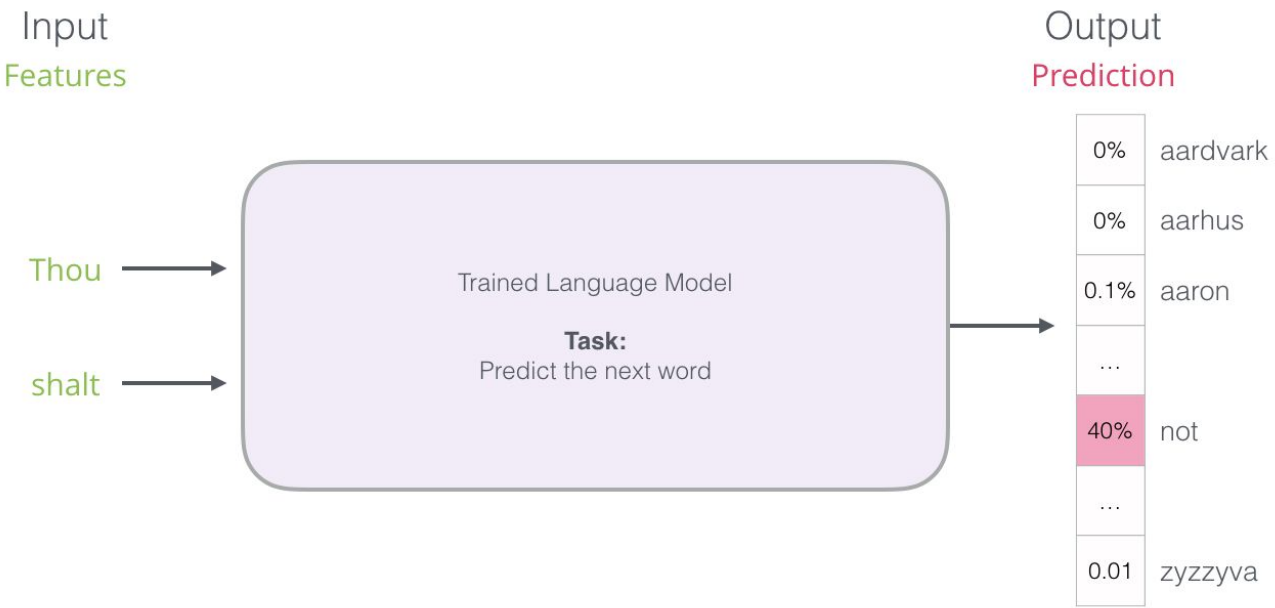
Традиционные методы - Bag of Words	Word Embeddings
• one hot encoding • Каждое слово в словаре представляется одной единицей в большом словаре • Информация в контексте не используется • Например если в словаре 10000 слов и слово Hello – 4-е слово в словаре, то оно представляется вектором 000100….000	• Представляет каждое слово как точку в пространстве с фиксированной размерностью • Unsupervised, строится на основе большого корпуса текста • К примеру слово Hello может быть представлено: [0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02]

Word2Vec: CBOW

- На вход принимаем one-hot encoding предыдущего слова
- Хотим предсказать индекс в словаре текущего слова
- Функция активации последнего слоя – softmax
- W или W’ - матрицы эмбеддингов
- Учим по неразмеченным данным, поэтому корпуса большие

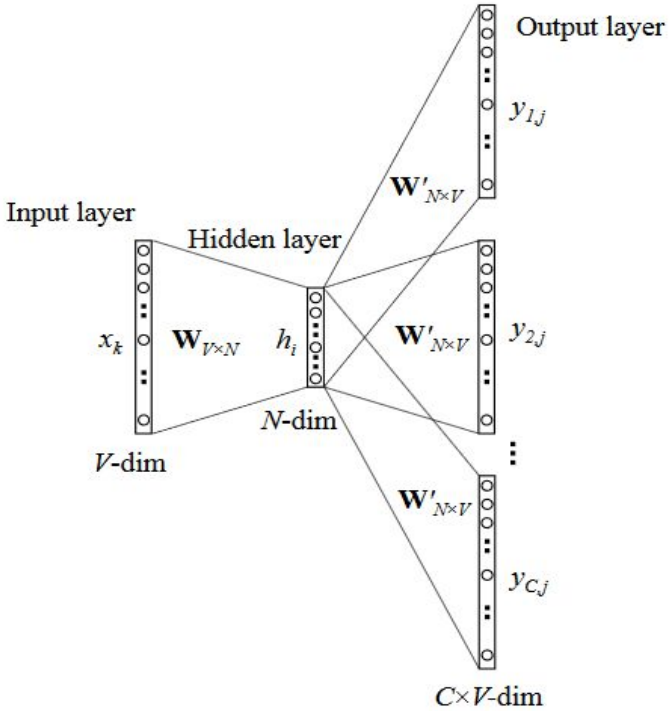


Word2Vec: CBOW



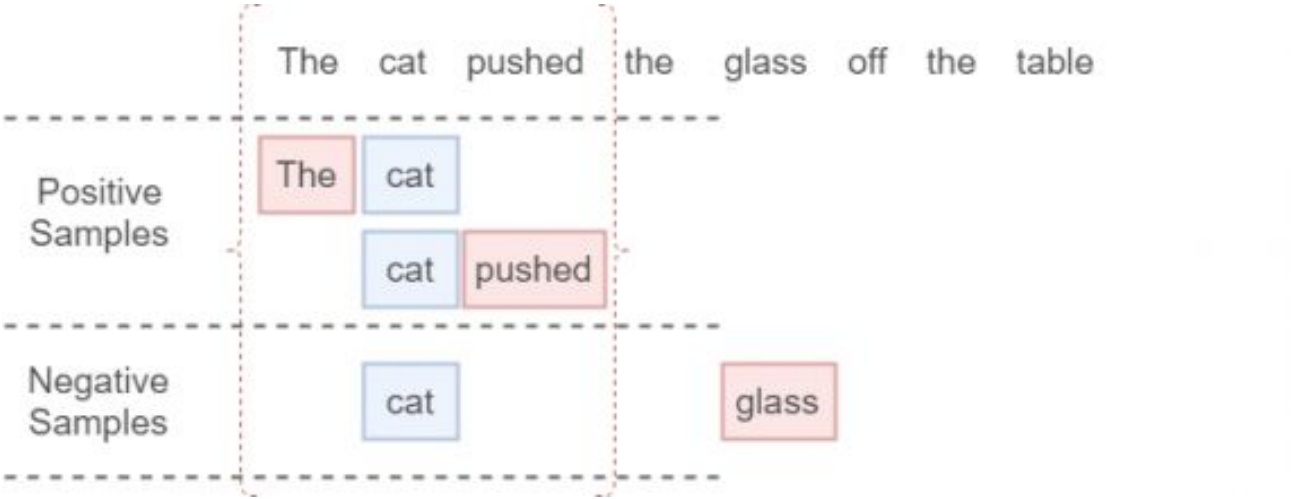
Word2Vec: SkipGram

- Предсказать контекст по слову
- Softmax в размерность словаря вычисляется очень долго
- Для оптимизации используют разные техники, но самая популярная – negative sampling

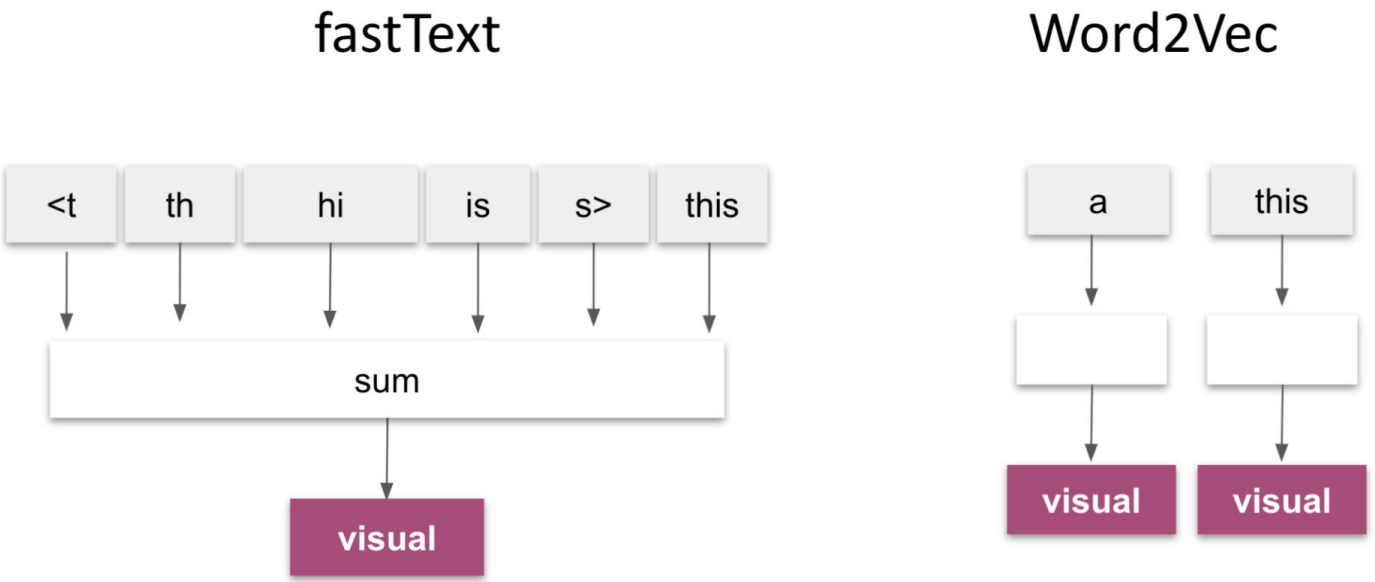


Negative Sampling

Размер словаря может вырасти до 100,000 значений или даже больше, что значительно усложняет вычисление softmax

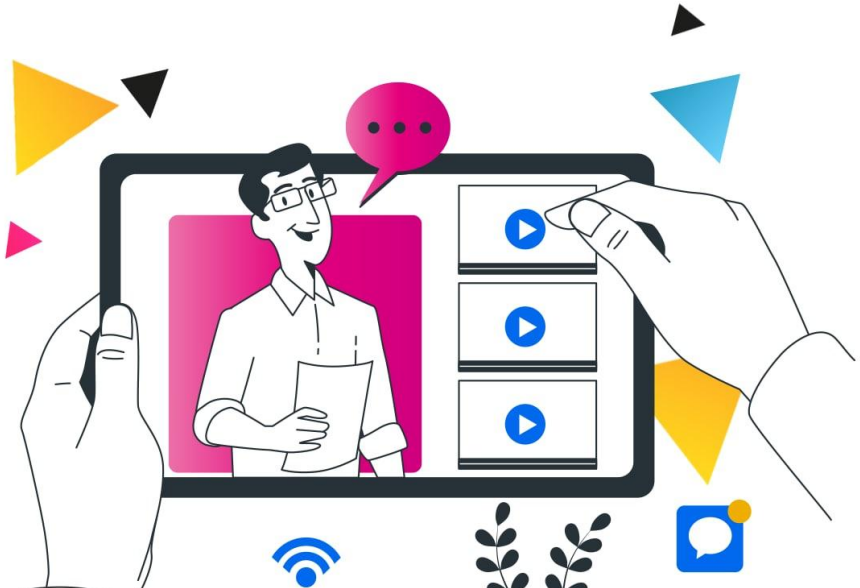


FastText


fastText
<t
th
hi
is
s>
this
sum
visual
Word2Vec
a
this
visual
visual

Практика (gensim)

Ваши вопросы?



Итоги занятия

Итоги занятия

1. Изучили методы предобработки текста с помощью pymorphy и nltk
2. Обучили модели CountVectorizer и TfIdfVectorizer
3. Поработали с моделью Word2Vec
4. Решили задачу классификации текста с помощью предобученной модели FastText

Дополнительные материалы

1. Метрики в torch https://torchmetrics.readthedocs.io/en/v0.8.2/index.html
2. Предобученные модели https://rusvectores.org/ru/models/

МФТИ

Пожалуйста, оставьте свой отзыв о семинаре

До встречи!