СБЕР УНИВЕРСИТЕТ

Программа Перезапуск
Модуль ML.
Занятие 2: NLP in prod

Преподаватель: Марат Гарафутдинов

Закрываем долги

Text processing: tools

- **NLTK**: stemming, lemmatizer
- BeautifulSoup (wotking with HTML)
- RegEx (re)
- PyMorphy2
- DIY instruments

Bag-of-Words (BoW)

- Для начала составляется словарь всех или наиболее часто употребляемых слов исходного датасета.
- Затем каждому тексту ставится в соответствие вектор длины словаря, где на i-ой позиции записывается количество вхождений i-ого слова.
- Такой подход уже позволяет сравнивать тексты, например при помощи косинусной меры.

- Однако у него есть множество недостатков:
- теряется информация о порядке слов;
- вектора представлений слишком большие и разреженные;
- вектора представлений не нормализованы.

Bag-of-Words (BoW)

TF IDF

Term frequency (TF)

- $\text{tf}(t, d)$ – frequency for term (or n-gram) t in document d
- Variants:

weighting scheme	TF weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$1 + \log(f_{t,d})$

TF IDF

Sentence A: The car is driven on the road.
Sentence B: The truck is driven on the highway.

(each sentence is a separate document)

Word	TF		IDF	TF * IDF	
	A	B		A	B
The	1/7	1/7			
Car	1/7	0			
Truck	0	1/7			
Is	1/7	1/7			
Driven	1/7	1/7			
On	1/7	1/7			
The	1/7	1/7			
Road	1/7	0			
Highway	0	1/7			

TF IDF

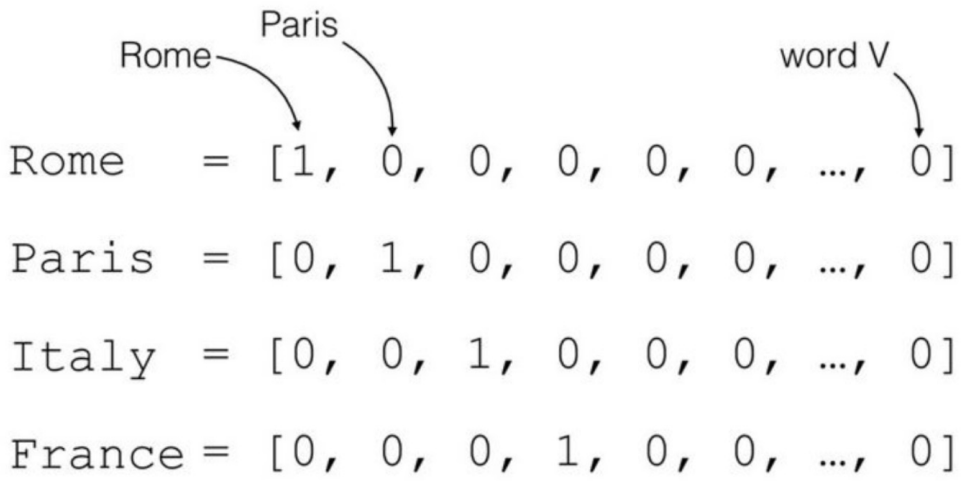
Sentence A: The car is driven on the road.
Sentence B: The truck is driven on the highway.

(each sentence is a separate document)

Word	TF		IDF	TF * IDF	
	A	B		A	B
The	1/7	1/7	log(2/2)=0	0	0
Car	1/7	0	log(2/1)=0.3	0.043	0
Truck	0	1/7	log(2/1)=0.3	0	0.043
Is	1/7	1/7	log(2/2)=0	0	0
Driven	1/7	1/7	log(2/2)=0	0	0
On	1/7	1/7	log(2/2)=0	0	0
The	1/7	1/7	log(2/2)=0	0	0
Road	1/7	0	log(2/1)=0.3	0.043	0
Highway	0	1/7	log(2/1)=0.3	0	0.043

Word Embeddings

- **One-hot vectors:**



Problems:

- Huge vectors
- VERY sparse
- No semantics or word similarity information included