
INTRODUCTION TO NLP

План лекции

1. Что такое текст
2. Задачи в NLP
3. Как обрабатывается текст
4. Препроцессинг текста
5. N-граммы
6. Bag of Words (Частотный словарь)
7. TF-IDF
8. Word2Vec
9. Doc2Vec

Что такое текст?

Текст - в общем плане связная и полная последовательность символов.

Задачи в NLP

1. Классификация
2. Регрессия
3. Перевод
4. Token classification
5. POS-tagging
6. QA
7. text2text generation
8. Retrieval

Предобработка текста

1. Удаление знаков препинания
2. Удаление стоп слов
3. Удаление ссылок
4. Удаление эмодзи
5. Приведение к нижнему регистру
6. Регулярные выражения
7. Удаление выбросов

Токенизация

Процесс разделения письменного языка на предложения-компоненты

Токенизация по предложениям [“Мама мыла раму”, “Папа изучал NLP”]

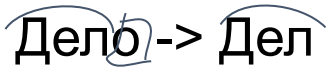
Токенизация по словамРазделение предложекний на слова [“Мама”, “мыла”, “раму”]

Токенизация по частям слов Разделение слов по морфемам [“Мам”, “а”, “мыл”, “а”, “раму”]

Токенизация по буквам
[“М”, “а”, “м”, “а”, “ ”, “м”, “ы”, “л”, “а”, “ ”, “р”, “а”, “м”, “у”]

Стемминг

Стемминг – это грубый эвристический процесс, который отрезает «лишнее» от корня слов, часто это приводит к потере словообразовательных суффиксов



Лемматизация

Лемматизация – это более тонкий процесс, который использует словарь и морфологический анализ, чтобы в итоге привести слово к его канонической форме – лемме.

Лемматизация: Сделал -> делать

Стемминг: Сделал -> дел

N-Граммы

Последовательность из N элементов.
Целью создание N-грамм является определения вероятности той или иной последовательности элементов.

Bag of words

Подсчет количества каждого слова в последовательности

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

TF-IDF

У частотного скоринга есть проблема: слова с наибольшей частотностью имеют, соответственно, наибольшую оценку. В этихсловах может быть не так много информационного выигрыша для модели, как в менее частых словах. Один из способов исправить ситуацию – понижать оценку слова, которое часто встречается во всех схожих документах. Это называется TF-IDF.

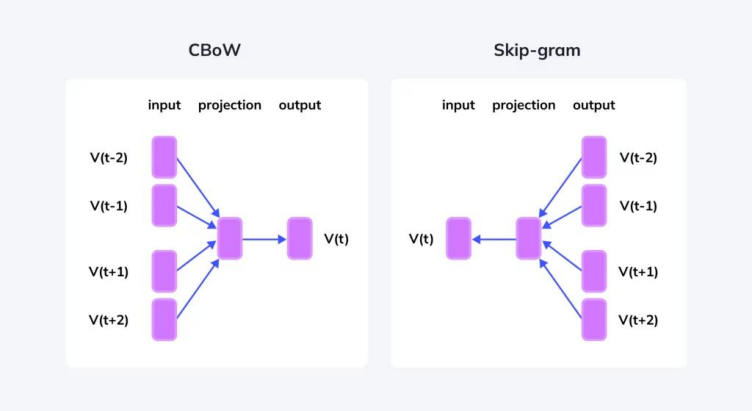
$W_{x,y} = TF_{x,y} * \log(N / DF_x)$

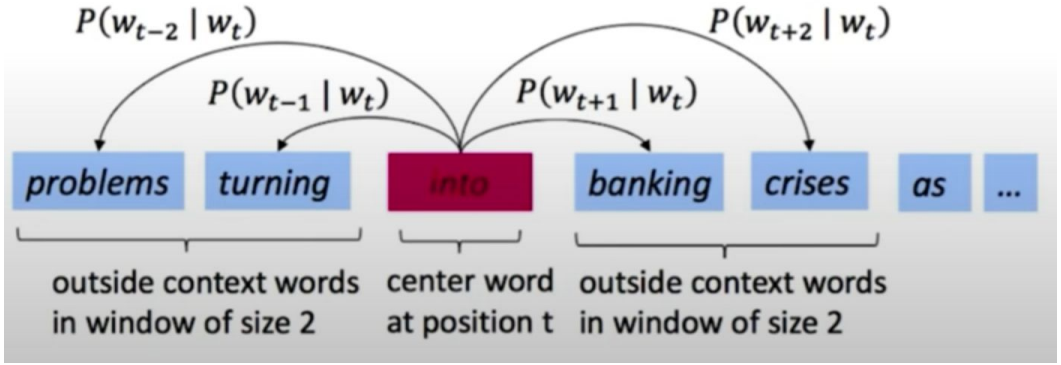
$TF_{x,y}$ - Частота слова x в сэмпле y
DF_x - Кол-во документов, содержащих x N - общее кол-во документов

Word2Vec, эмбеддинг

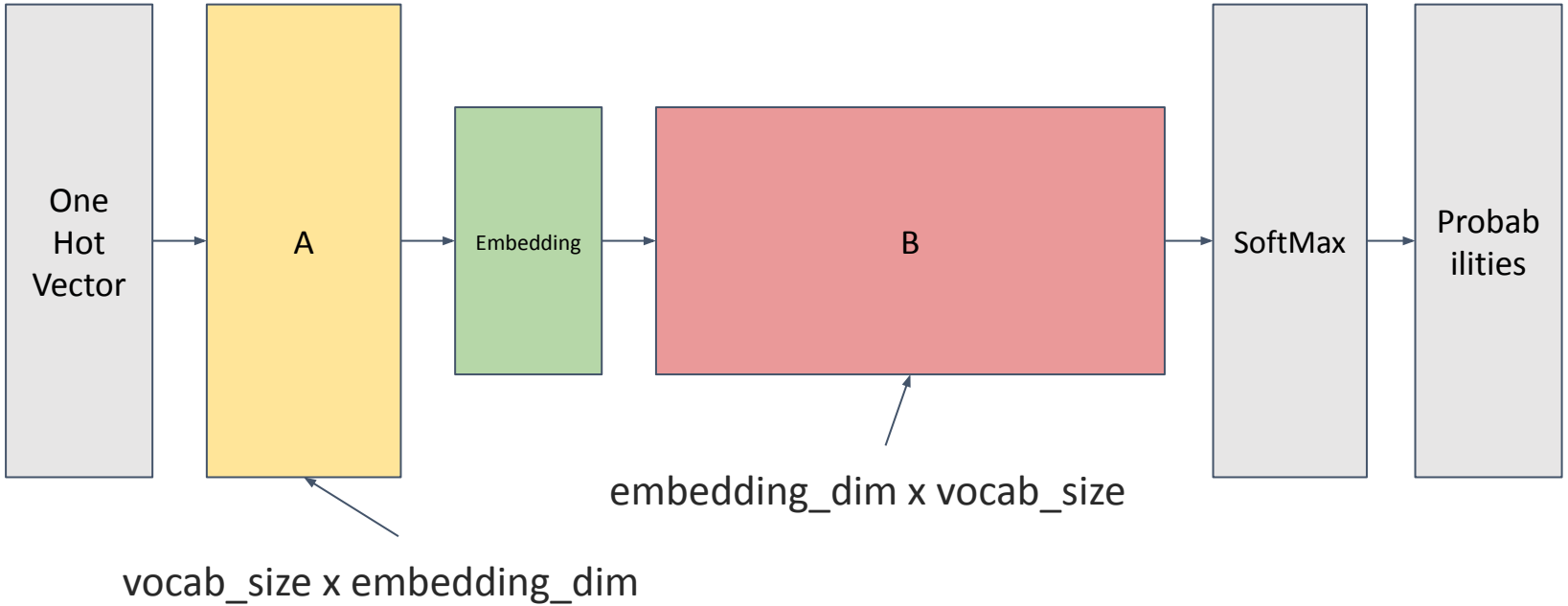
Однако Word2Vec - нейросеть для получения текстовых эмбеддингов

CBOW: Предсказать по словам из окна недостающее
Skip Gram: Предсказать по 1 слову остальные в предложении





Word2Vec архитектура



Loss

$$H(P^* \mid P) = -\sum_i \underbrace{P^*(i)}_{\text{TRUE CLASS DISTIRBUTION}} \log \underbrace{P(i)}_{\text{PREDICTED CLASS DISTIRBUTION}}$$

Свойства Word2Vec

1. Косинус между 2 эбеддингами показывает семантические сходство

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

2. Семантическое вычитание и сложение