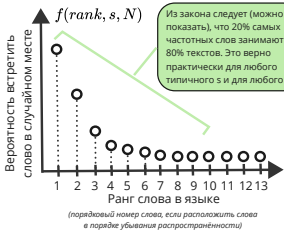
Занятие 7
Тексты

Определения

Корпус — набор текстов.
Документ — текст состоящий из одного или нескольких предложений.
Токен — минимальная единица языка, принимаемая во внимание при анализе. Иногда это может быть слово, (иногда часть слова, иногда отдельный символ), знак препинания, цифра и тд.
Эмбеддинг — вектор (набор чисел), которым можно представить слово/токен/предложение.

Закон Ципфа

Закон Ципфа утверждает, что в языке, состоящим из N уникальных слов, вероятность встретить слово ранга rank в случайном месте равняется:

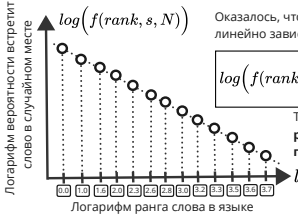


$$f(rank, s, N) = \frac{1/rank^s}{\sum_{n=1}^{N} 1/n^s}$$

N — количество слов в словаре данного языка.

s — коэффициент степенного закона, показывающий скорость убывания частоты встречаемости слова. Чем больше s, тем реже редкие слова относительно популярных слов. (тем легче набирать словарный запас) Типичные значения 0.95 — 1.2.

$rank$ — порядковый номер слова в отсортированном по убыванию частотности списке слов словаря



Оказалось, что логарифм частоты встречаемости слова линейно зависит от логарифма его ранга:

$$log\Big(f(rank, s, k)\Big) = -log\Big(\sum_{n=1}^{N} 1/n^s\Big) - s \cdot log(rank)$$

Таким образом, **коэффициент s оказывается равным тангенсу угла наклона построенной прямой**.

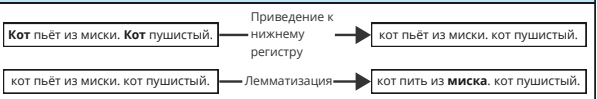
Совершенно логично, что в голове распределения (маленький ранг) находятся слова, не влияющие на смысл текстов (предлоги, союзы, частицы). В хвосте наоборот находятся слишком редкие слова. При анализе текстов для простоты вычислений обычно выбирают часть распределения Ципфа, удаляя голову и хвост.

	верхний график		*нижний график*	
word	rank	freq	log(rank)	log(freq)
the	1	0.071	0.0	-3.81
of	2	0.037	1.0	-4.76
and	3	0.029	1.6	-5.11
...	...	...	...	...
like	78	0.001	6.3	-9.97
...	...	...	...	...
education	426	0.0002	8.7	-12.29
...	...	...	...	...
discover	2681	0.0000	11.4	-16.6
...	...	...	...	...

Таблица: частота встречаемости некоторых слов в английском языке, построенная на основе Брауновского корпуса. (Корпус = набор текстов). Корпус был создан в 1970-е в университете Брауна и представляет собой набор из 500 текстов, в сумме содержащих около 1 миллиона уникальных слов.

Занятие 7
Тексты

Предобработка текстов



Векторизация текстов

BOW (Bag Of Words)

Чтобы работать с текстами (проводить классификацию/кластеризация/регрессию), необходимо представить тексты в виде набора чисел.

Признаки текстов корпуса (фичи)

Корпус текстов (Набор текстов)	бежать	за	из	кот	лежать	миска	на	пить	пушистый	собака	солнышко
Кот пьёт из миски. Кот пушистый.	0	0	1	2	0	1	0	1	1	0	0
Собака бежит за котом.	1	1	0	1	0	0	0	0	0	1	0
Собака лежит на солнышке.	0	0	0	0	1	0	1	0	0	1	1

Вектор предложения (эмбеддинг) "Собака лежит на солнышке"

Метод реализован в sklearn.CountVectorizer()

TF-IDF

Оказалось, что вектора BOW получаются слишком зашумлёнными популярными словами типа предлогов, союзов и местоимений.

$$TF(word, document) = \frac{Количество\ повторений\ данного\ слова\ в\ документе}{Общее\ количество\ слов\ в\ документе}$$

Чем больше TF, тем чаще данное слово встречается в данном тексте.

$$DF(word) = \frac{Количество\ документов\ с\ данным\ словом}{Общее\ количество\ документов}$$

Чем больше DF, тем чаще данное слово встречается в корпусе (наборе текстов).

$$IDF(word) = log\Big(\frac{Общее\ количество\ документов}{Количество\ документов\ с\ данным\ словом}\Big) + 1$$

Чем больше IDF, тем в меньшем количестве текстов встречается данное слово. Величина 1/DF логарифмируется, чтобы уменьшить очень большие значения. Единица прибавляется, чтобы для слов, которые встречаются во всех документах, не обнулялось произведение TF*IDF:

$$TFIDF(word, document) = TF(word, document) \cdot IDF(word)$$

	бежать	за	из	кот	лежать	миска	на	пить	пушистый	собака	солнышко
Кот пьёт из миски. Кот пушистый.	0	0	1.69	2.58	0	1.69	0	1.69	1.69	0	0
Собака бежит за котом.	1.69	1.69	0	1.29	0	0	0	0	0	1	0
Собака лежит на солнышке.	0	0	0	0	1.69	0	1.69	0	0	1.29	1.69

Метод реализован в sklearn.TfidfVectorizer()

Сходство предложений: с помощью скалярного произведения эмебдингов TF-IDF можно определять смысловое сходство двух текстов.

Занятие 7

Тексты

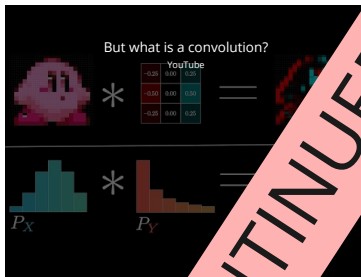
Векторизация изображений

Разворачиваем двумерную картинку в единый вектор. Значения пикселей — это фичи для будущего классификатора.

Картинка 4x4 пикселя

Количество признаков 1x16

Свёртки



TO BE CONTINUED