Векторное представление слов

Word embeddings. Introduction

Текст и МО

- Алгоритмы МО работают с числовыми данными (признаками)
- А мы работаем с текстами и словами
- Как превратить текст или слова в цифры?

P.S. до того, как превращать текст в чиселки, нам нужно сделать **предобработку**: нормализовать, токенизировать, лемматизировать, решить момент с пунктуацией

One-hot encoding

- Токенизируем текст
- делаем множество из слов (токены или леммы?)
- каждому слову на месте его «номера» ставим 1, остальное заполняем нулями.
- Получаем матрицу с векторами слов. Как называется такая матрица?

Плюсы? Минусы?

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Мешок слов

aka bag-of-words, векторное представление текста

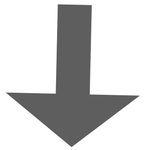
- У нас есть несколько документов (напр., письма со спамом/не спамом)
- делаем множество из слов во всех документах
- считаем и заносим в вектор число, сколько раз каждое слово встретилось в каждом документе.

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Мешок слов

aka bag-of-words

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1



the cat sat: `[1, 1, 1, 0, 0, 0]`

the cat sat in the hat: `[2, 1, 1, 1, 1, 0]`

the cat with the hat: `[2, 1, 0, 0, 1, 1]`

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List

for
is
of
the
to

Мешок слов

aka bag-of-words

- Упрощенное векторное представление текста
- У нас есть несколько документов (напр., письма со спамом/не спамом), считаем и заносим в вектор число, сколько раз каждое слово встретилось в каждом документе.
- Можно считать по N-граммам

- Плюсы? Минусы?
- *В каких задачах можно использовать?

Задание

Посчитайте мешок слов для этих 3 текстов

good movie
not a good movie
did not like

Задание

Посчитайте мешок слов для этих 3 текстов

good movie
not a good movie
did not like

good	movie	not	a	did	like

Задание

- сначала делаем множество слов
- считаем кол-во каждого слова в каждом документе

good movie
not a good movie
did not like

good	movie	not	a	did	like
1	1	0	0	0	0
1	1	1	1	0	0
0	0	1	0	1	1

Tf-idf

Мешок слов на стероидах

Tf-idf

Мешок слов на стероидах

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Берем «документы»

- d_1: "*The sky is blue.*
- d_2: "*The sun is bright today.*"
- d_3: "*The sun in the sky is bright.*"
- d_4: "*We can see the shining sun, the bright sun.*"

Выкидываем стоп-слова

- d_1: "*sky blue*
- d_2: "*sun bright today*"
- d_3: "*sun sky bright*"
- d_4: "*can see shining sun bright sun*"

Рассчитываем частоту слов (term frequency)

$$f_{t,d}$$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

$$\mathtt{tf}(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

Рассчитываем обратную частоту слов в документе (idf)

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0
n_t	**1**	**3**	**1**	**1**	**1**	**2**	**3**	**1**

$N = 4$

$$\texttt{idf}(t, D) = \log_{10} \frac{N}{n_t}$$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$\log_{10} \frac{4}{1} = 0.602$

$\log_{10} \frac{4}{3} = 0.125$

Получаем матрицу Tf-idf

$\mathtt{tf}(t, d)$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

x

$\mathtt{idf}(t, D)$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

- TF-IDF: Multiply TF and IDF scores, use to rank importance of words within documents
  - Most important word for each document is highlighted

$\mathtt{tfidf}(t, d, D) = \mathtt{tf}(t, d) \cdot \mathtt{idf}(t, D)$

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	**0.100**	0.0417	0
4	0	0.0209	**0.100**	**0.100**	**0.100**	0	0.0417	0

Tf-idf

Мешок слов на стероидах

- показывает важность (вес) слова в документе
- занижает вес слишком частотных слов

Задание

Посчитайте теперь TF-IDF для этих текстов

good movie
not a good movie
did not like

Задание

Посчитайте теперь TF-IDF для этих текстов

good movie
not a good movie
did not like

good	movie	not	a	did	like

Слова, а не тексты

- ОНЕ слишком плох
- Мешок слов и TF-IDF являютя представлениями текстов/документов, но не конкретных слов. Т.е. работают с классификациями текстов.
- А что если для задачи нужно переводить в векторы именно слова?

Word2vec

Words to vectors

- Вектор-представление для каждого слова
- Что мы хотим от такого вектора?

Word2vec

Words to vectors

- Вектор-представление для каждого слова
- Что мы хотим от такого вектора?

- Отражал смысл слова (похожие слова имели бы похожие векторы)

Лирическое отступление

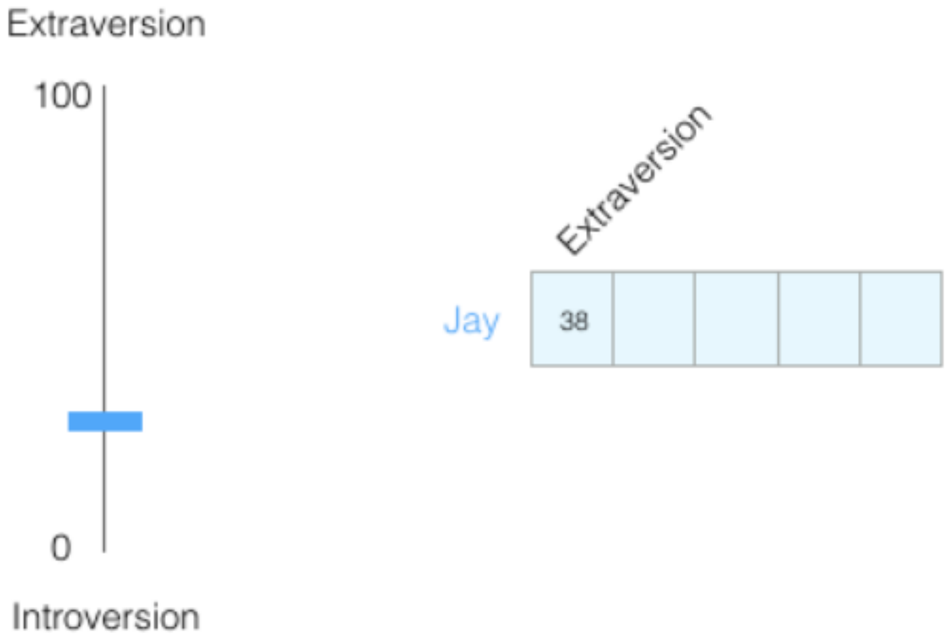
Психологический тест «Большая пятерка»

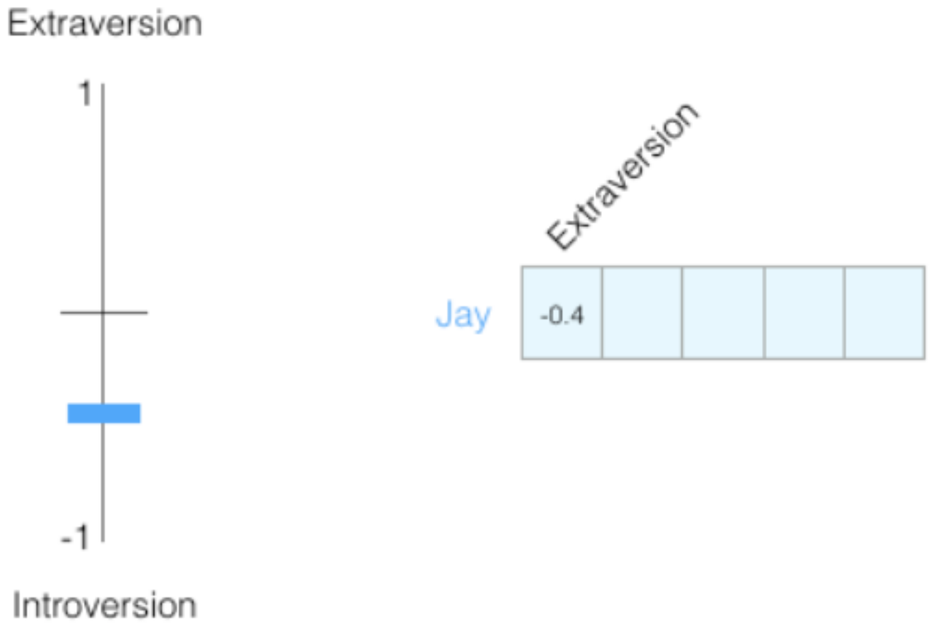
Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100

https://habr.com/ru/post/446530/

Большая пятерка (первая черта)

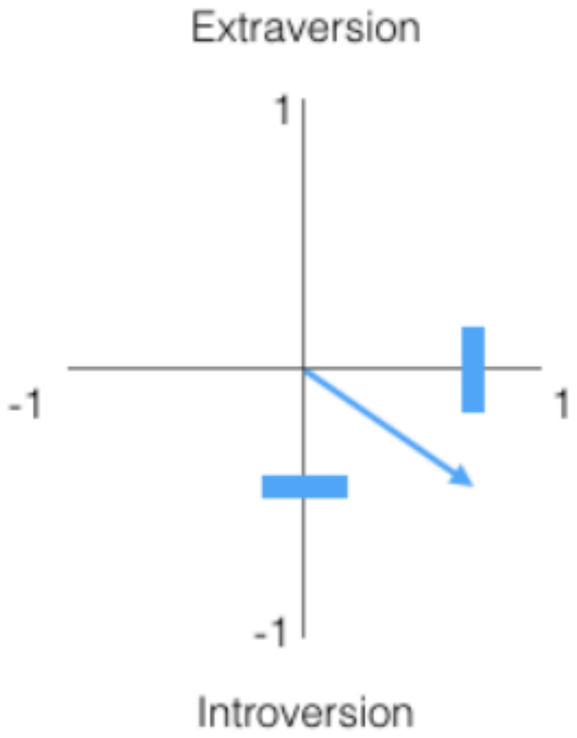
Масштабируем по шкале от 1 до -1

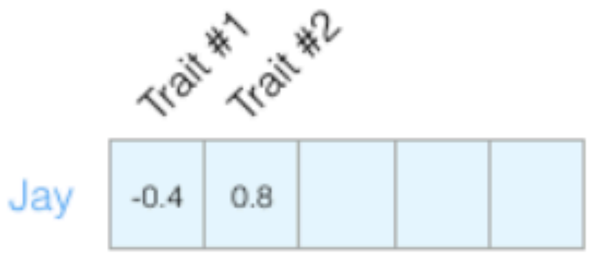




https://habr.com/ru/post/446530/

Большая пятерка (добавим еще черту)





https://habr.com/ru/post/446530/

Большая пятерка

Еще 2 человека ответили на 2 вопроса теста

Кто больше похож на синего?

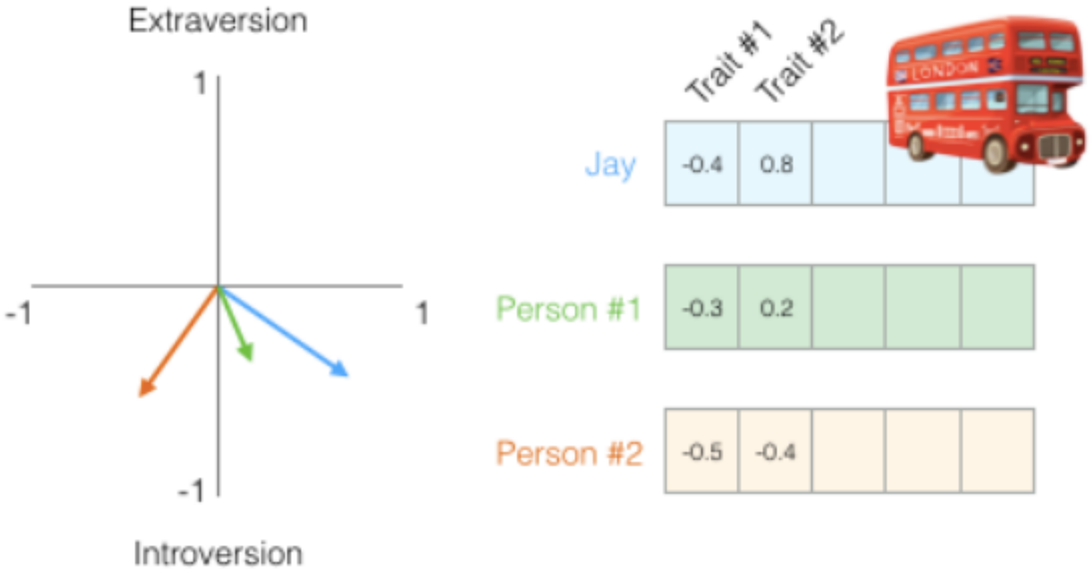
https://habr.com/ru/post/446530/

Большая пятерка

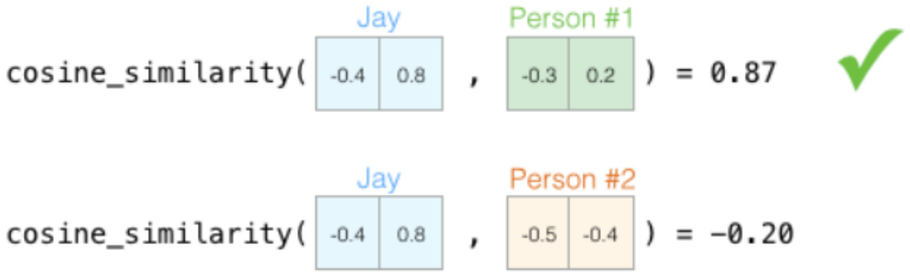
- В 3-мерном пространстве можно определить на глаз
- Но лучше не полагаться на это и пользоваться субъективным инструментом
- Есть идеи?

https://habr.com/ru/post/446530/

Большая пятерка



- В 3-мерном пространстве можно определить на глаз
- Но лучше не полагаться на это и пользоваться субъективным инструментом
- Косинусная близость (угол между векторами)

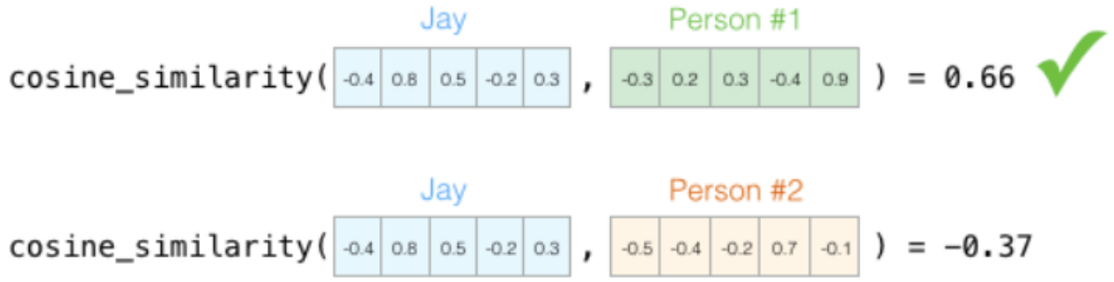


https://habr.com/ru/post/446530/

Многомерное пространство

	Trait #1	Trait #2	Trait #3	Trait #4	Trait #5
Jay	-0.4	0.8	0.5	-0.2	0.3
Person #1	-0.3	0.2	0.3	-0.4	0.9
Person #2	-0.5	-0.4	-0.2	0.7	-0.1

- Больше 3-мерного пространства не можем визуализировать
- Теперь точно сравниваем только с помощью косинусной близости



https://habr.com/ru/post/446530/

Переходим к словам

Почти word2vec

Дистрибутивная гипотеза: смысл слова - в его контексте, то есть в словах, с которым оно чаще всего встречается.

Откуда взять цифры для векторов всех слов?

Переходим к словам

Почти word2vec

Дистрибутивная гипотеза: смысл слова - в его контексте, то есть в словах, с которым оно чаще всего встречается.

Откуда взять цифры для векторов всех слов?

	редис	картошка	кот	...	собака
редис	-	5	1		0
картошка	5	-	0		1
кот	1	0	-		6
...					
собака	0	1	6		-

Эмбеддинг

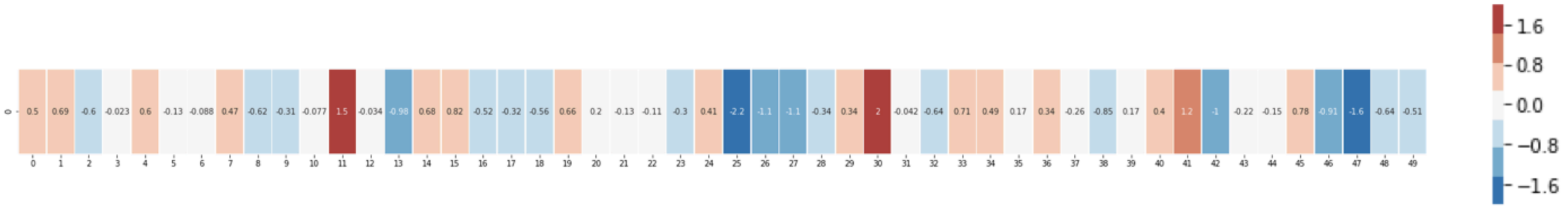
Векторное представление слова

Вектор слова «король»

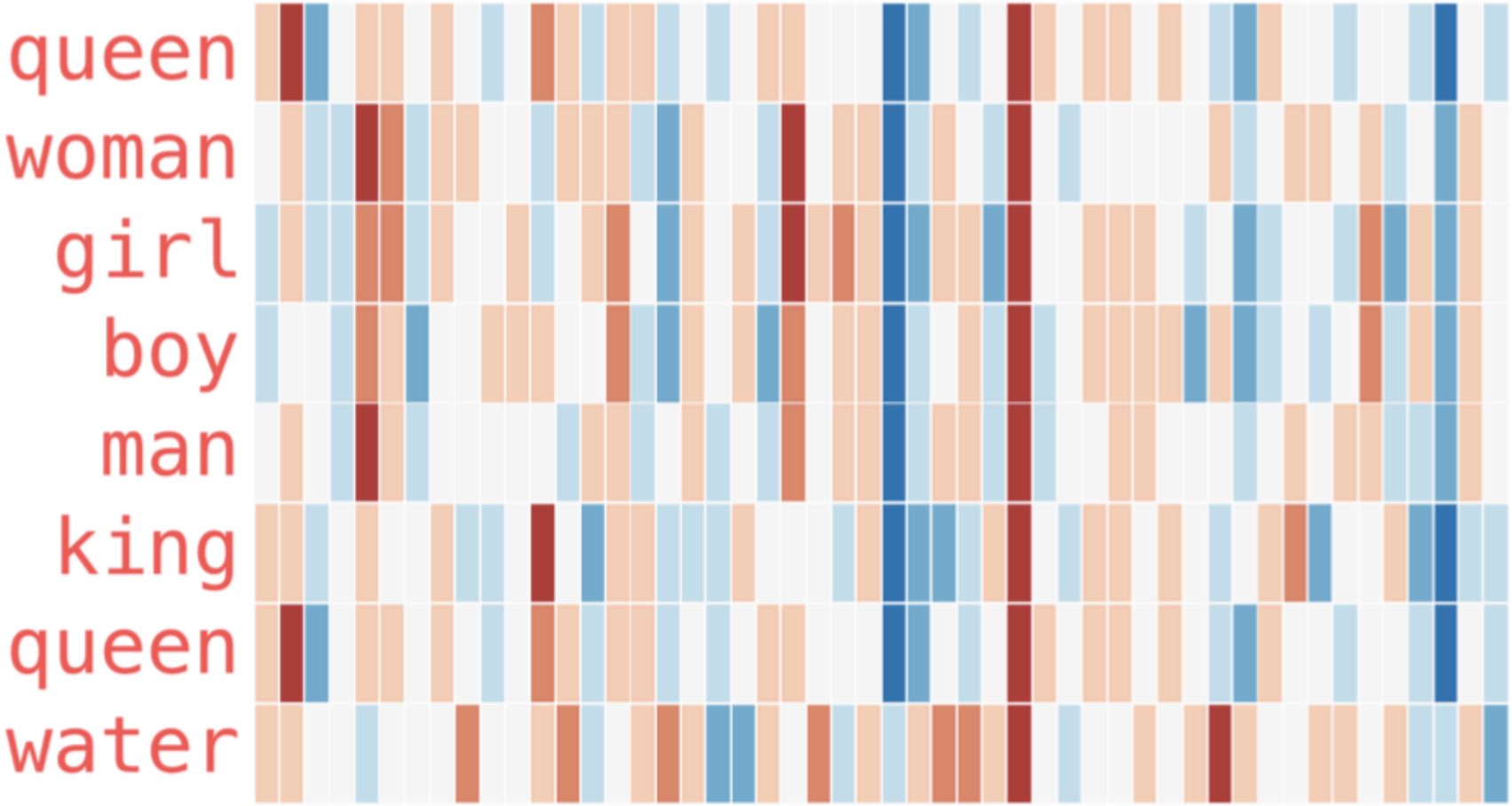
```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 ,
0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 ,
0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 ,
-0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 ,
-0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 ,
0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137
, -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

Эмбеддинг «Король»

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 ,
0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 ,
0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 ,
-0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 ,
-0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 ,
0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137
, -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

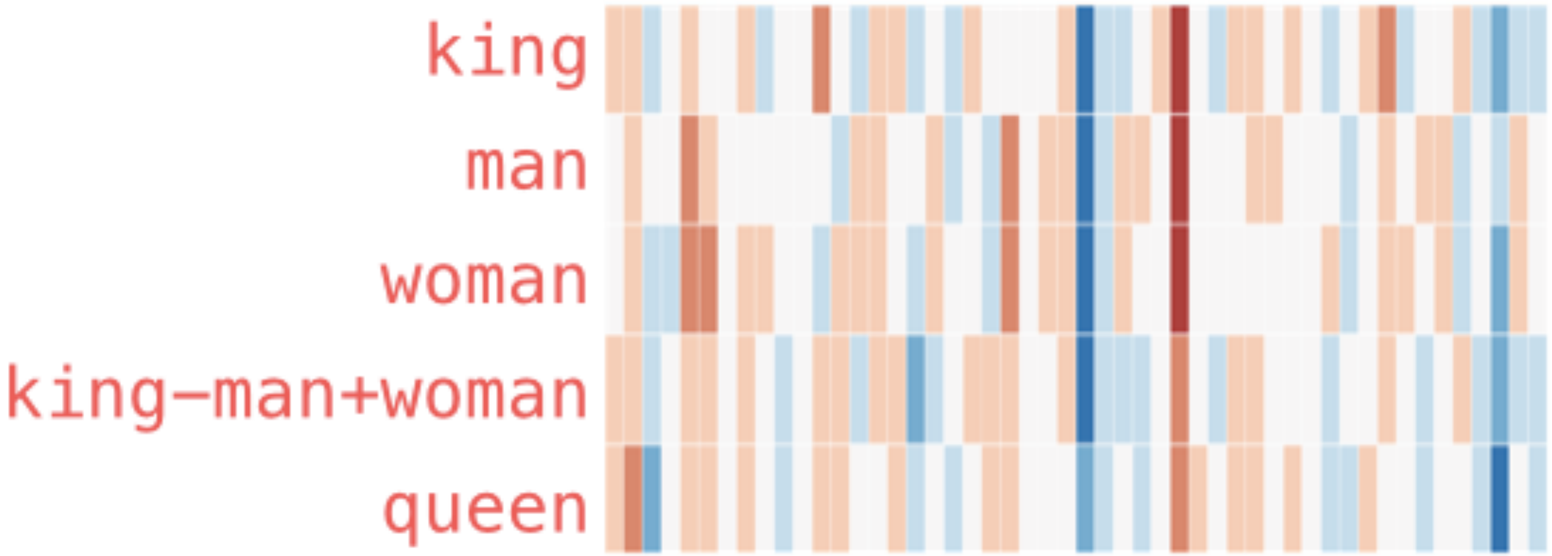


Сравните эмбеддинги



Эмбеддинги

king − man + woman ~= queen



Полученный вектор от вычисления «король−мужчина+женщина» не совсем равен «королеве», но это наиболее близкий результат из 400 000 вложений слов в наборе данных

Полезные материалы

1. https://habr.com/ru/post/446530/
2. https://sysblok.ru/knowhow/word2vec-pokazhi-mne-svoj-kontekst-i-ja-skazhu-kto-ty/
3. https://rusvectores.org/ru/

Блокнот с word2vec и теорией:

- https://github.com/hse-ds/iad-deep-learning/blob/master/2021/seminars/sem07/sem07_solution.ipynb