СПЗ - 3

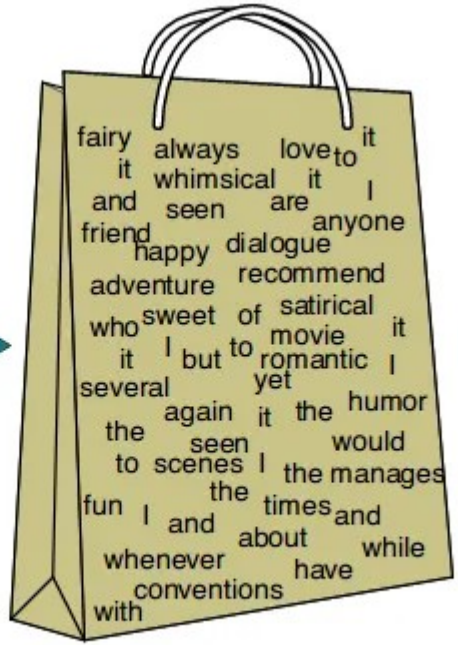
Обработка текста

- Предобработка текста
- Векторизация текста
  - Bag of Words
  - Bag of N-gramms
  - TF-IDF
- Word Embeddings

- Предобработка текста
  - Токенизация
  - Нормалицация
    - Стемминг
      - Potter
    - Лемматизация

Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



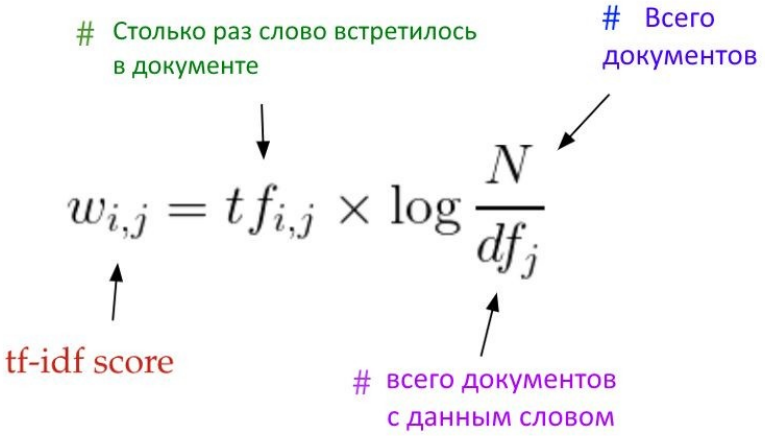
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Bag of Words

	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

- A = “The car is driven on the road”;
- B = “The truck is driven on the highway”



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

- Преобразовать тексты к числовому представлению
  - Классическими и простыми средствами для этого являются bag-of-words и tf-idf
- В sklearn есть удобные классы-векторайзеры, реализующие bag-of-words и tf-idf (и трансформер для преобразования от bag-of-words к tf-idf)
- Они реализуют токенизацию, векторизацию и отсечение стоп-слов
- У векторайзеров много параметров, позволяющих добиваться наиболее удачного представления текста. Знание этих параметров (и понимание того, как они работают) может сильно повысить эффективность обучения и качество результата
- Отдельно токенизацию, стемминг и прочее можно делать, например, средствами NLTK