Modeling & Graphing Text

Unstructured Text

The happy fox jumped over the sad fox

The angry dog ran at the upset foxes

The sad fox ran away from the mean dog

The lazy cat napped on my lap

The lion napped on the hill

The cat jumped on the happy fox

Structured Data

sent_id	word
1	happy
1	fox
1	jumped
1	sad
1	fox
2	angry
2	dog
2	ran
2	upset
2	foxes

Removing Stop Words

The happy fox jumped over the sad fox

The angry dog ran at the upset foxes

The sad fox ran away from the mean dog

The lazy cat napped on my lap

The lion napped on the hill

The cat jumped on the happy fox

Document Term Matrix

	angry	cat	dog	fox	foxes	happy	hill	jumped
1	0	0	0	2	0	1	0	1
2	1	0	1	0	1	0	0	0
3	0	0	1	1	0	0	0	0
4	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0
6	0	1	0	1	0	1	0	1

Term Frequencies

The **fox** jumped over the other **fox**

~~The dog ran at the foxes~~

~~The fox ran away from the dog~~

~~The cat napped on my lap~~

~~The lion napped on the hill~~

~~The cat jumped on the fox~~

Inverse Document Frequencies

The fox jumped over the other fox

The dog ran at the foxes

The fox ran away from the dog

The cat napped on my lap

The lion napped on the hill

The cat jumped on the fox

Inverse Document Frequencies

The fox jumped over the other fox

The dog ran at the foxes

The fox ran away from the dog

The cat napped on my lap

The lion napped on the hill

The cat jumped on the fox

TF-IDF

sent_id	word	n	tf	idf	tf_idf
5	hill	1	0.3333333	1.7917595	0.5972532
5	lion	1	0.3333333	1.7917595	0.5972532
4	lap	1	0.2500000	1.7917595	0.4479399
4	lazy	1	0.2500000	1.7917595	0.4479399
5	napped	1	0.3333333	1.0986123	0.3662041
2	angry	1	0.2000000	1.7917595	0.3583519
2	foxes	1	0.2000000	1.7917595	0.3583519
2	upset	1	0.2000000	1.7917595	0.3583519
1	fox	2	0.4000000	0.6931472	0.2772589
3	dog	1	0.2500000	1.0986123	0.2746531
3	ran	1	0.2500000	1.0986123	0.2746531
3	sad	1	0.2500000	1.0986123	0.2746531

Cosine Similarity: What is similar to Sentence 1?

item1	item2	similarity
1	3	0.4546847
1	6	0.7078296

The happy fox jumped over the sad fox

The angry dog ran at the upset foxes

The sad fox ran away from the mean dog

The lazy cat napped on my lap

The lion napped on the hill

The cat jumped on the happy fox

Unsupervised Topic Modelling: K =3

	angry	cat	dog	fox	foxes	happy	hill	jumped	lap	lazy	lion	napped	ran
1	0	0	0	2	0	1	0	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0	0	0	0	1
3	0	0	1	1	0	0	0	0	0	0	0	0	1
4	0	1	0	0	0	0	0	0	1	1	0	1	0
5	0	0	0	0	0	0	1	0	0	0	1	1	0
6	0	1	0	1	0	1	0	1	0	0	0	0	0

Unsupervised Topic Modelling: K = 3

	angry	cat	dog	fox	foxes	happy	hill	jumped	lap	lazy	lion	napped	ran
1	0	0	0	2	0	1	0	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0	0	0	0	1
3	0	0	1	1	0	0	0	0	0	0	0	0	1
4	0	1	0	0	0	0	0	0	1	1	0	1	0
5	0	0	0	0	0	0	1	0	0	0	1	1	0
6	0	1	0	1	0	1	0	1	0	0	0	0	0

Color the trolls accordng to their topic