Data Representation and Exploration

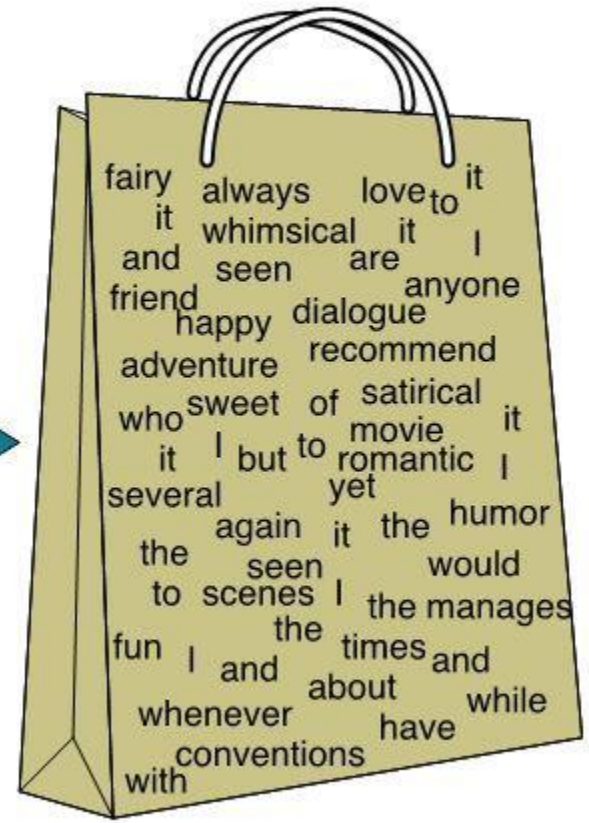
AI4ALL Princeton: NLP Group

How do we represent a body of text?

How do we represent many words?

The Bag of Words Representation

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
…	…

http://thecaffeinedev.com/twitter-sentiment-analysis-intro-machine-learning-beginners-part2/

Some words are more important than others

Some words are more important than others

How can we use this to improve our representation?

Term Frequency

$$\frac{word\ frequency\ in\ document}{length\ of\ document}$$

Inverse Document Frequency

$$log \frac{number\ of\ documents}{number\ of\ documents\ with\ word}$$

Term Frequency Inverse Document Frequency (TF-IDF)

$$\frac{word\ frequency\ in\ document}{length\ of\ document} * log\frac{number\ of\ documents}{number\ of\ documents\ with\ word}$$

tf-idf selects informative terms

DC-9 WITH 55 ABOARD CRASHES; AT LEAST 16 DEAD

CHARLOTTE, NC, (Reuter)

A USAir DC-9 with 55 people on board crashed and burst into flames during a thunderstorm after missing an approach to Charlotte's international airport Saturday, killing at least 16 people. The flight, which originated in Columbia, South Carolina and was on its final approach, hit a house near the airport runway and caught fire, said Jerry Orr, aviation director at Charlotte-Douglas International Airport. Orr said 16 people were dead, six were missing and presumed dead and 33 were taken to local hospitals. USAir reported 18 dead. Rescue teams fought to save lives inside the wreckage of the plane, which split into three sections on impact at about 6:50 p.m. EDT as the plane was trying to land at Charlotte during heavy storms.

...

top 15 terms ranked by

frequency		highest idf		tf * idf	
32	the	1.00	tdt000077	3.20	orr
16	were	1.00	picknickers	2.81	charlotte
14	said	0.93	sreaming	2.65	payne
12	and	0.93	timmy	2.48	dc
12	to	0.86	6thld	2.24	usair
11	a	0.80	orr	2.00	plane
10	of	0.78	1016	1.93	crash
9	at	0.76	bergen	1.74	bones
9	was	0.75	dripping	1.63	survivors
7	in	0.73	abrams	1.50	dripping
6	on	0.72	0419	1.49	wreckage
6	they	0.69	fuselage	1.35	dead
6	people	0.66	nc	1.29	hospitals
6	had	0.66	thunderstorm	1.27	airport
6	plane	0.66	payne	1.23	55

https://www.youtube.com/watch?v=zvFGNpbAfEI

How can we improve representations?

Princeton AI4ALL is intended to increase diversity and inclusion in the field of artificial intelligence.

http://ai4all.princeton.edu/about

Princeton AI4ALL is intended to increase diversity and inclusion in the field of artificial intelligence.

['princeton', 'ai4al', 'is', 'intend', 'to', 'increas', 'divers', 'and', 'inclus', 'in', 'the', 'field', 'of', 'artifici', 'intellig', '.']

http://ai4all.princeton.edu/about

Exploring a new dataset

What can we explore for a new dataset?

- Statistics about the data?
- Visualizations?
- Manual analysis?

What can we explore for a new dataset?

- Statistics about the data?
- Visualizations?
- Manual analysis?

Discuss with your neighbors