CS 205

Text Similarity

Lecture #13 | Oct. 4, 2016

L_1: So it's gonna be forever
L_2: Or it's gonna go down in flames
L_3: You can tell me when it's over

- Taylor Swift, Blank Space

Text Similarity

tf-idf is a classical measure of text similarity, allowing us to establish a base similarity metric between similar regions of text.

tf, **t**erm **f**requency asks "how often does the word appear?"
idf, **i**nverse **d**ocument **f**requency asks "how rare is the world?"

Calculating tf-idf

First, we calculate the tf-idf for every word in every line:

(a): How many regions does our document have? $\#_{\text{regions}}$:

(b): How many regions does [**it's**] appear? $\#(\text{it's})_{\text{regions}}$:

(c): How many times does [**it's**] appear in L_1? **tf** = L_1(it's):

(d): Calculate the idf of [**it's**] on L_1: **idf** = log((a) / (b)):

(e): Calculate tf-idf of [**it's**] on L_1: **tf-idf** = (c) * (d):

Second, we calculate the **cosine similarity** of each tf-idf:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

How often does [**it's**] appear in L1?

			tf			**idf**			**tf-idf**		
	#regions	**#regions(word)**	**L_1**	**L_2**	**L_3**	**L_1**	**L_2**	**L_3**	**L_1**	**L_2**	**L_3**
so											
it's											
gonna											
be											
forever											
or											
go											
down											
in											
flames											
you											
can											
tell											
me											
when											
over											

How do we set up a cosine similarity?