27 Mar 2023

Goal: compute how close two pieces of text are

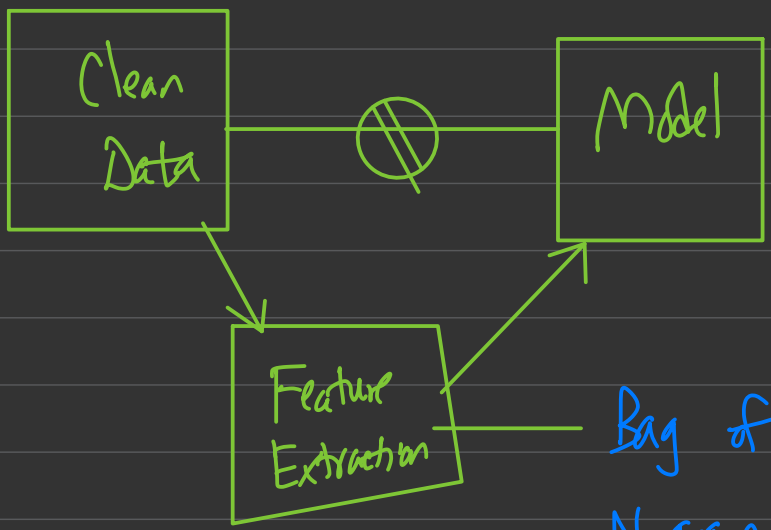
- (semantic similarity) meaning
- surface closeness (lexical similarity)

The cat ate the mouse
1 2 3

The mouse ate the cat food.
1 2 3 4

Clustering, Redundancy removal, Information Retrieval

What is Feature Extraction?



- Bag of Words ✓
- N-gram
- Document Term Matrix
- TF-IDF ✓
- Word embedding ✓

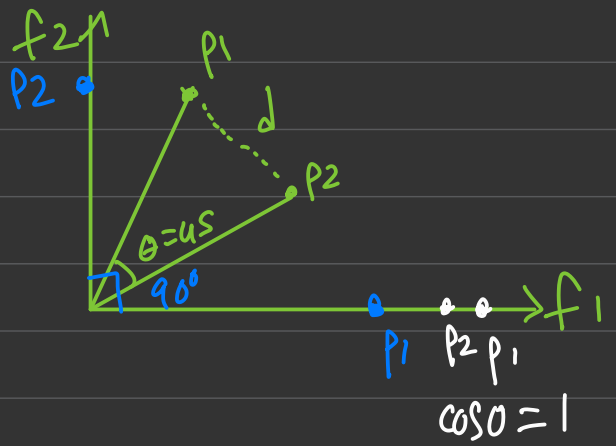
Cosine Similarity: Recommendation Systems

Similarity ↓ Distance ↑ [-1 to +1]

↑ ↓

cos_sim = cos θ

angle blw blw p1 & p2

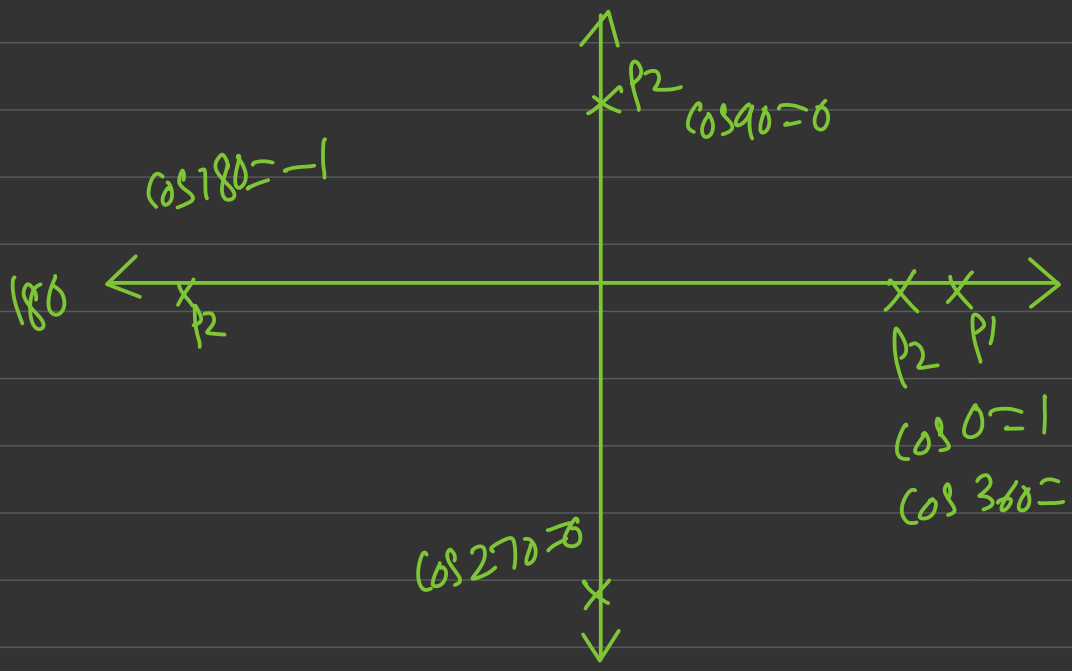


= cos 45

= 0.53

cos 0 = 1

cos 90 = 0



Count Vectorizer: (Bag of Words):

S1: I love playing
S2: We love to play cricket
S3: Playing is fun

→ Text Preprocessing (lower, rem. punct, rem. stopwords, stemming/Lemmatization)

S1: love play
S2: love play cricket
S3: play fun

Unique words = {love, play, cricket, fun}

	love	play	cricket	fun
1	1	1	0	0
2	1	1	1	0
3	0	1	0	1
Σ	2	3	1	1

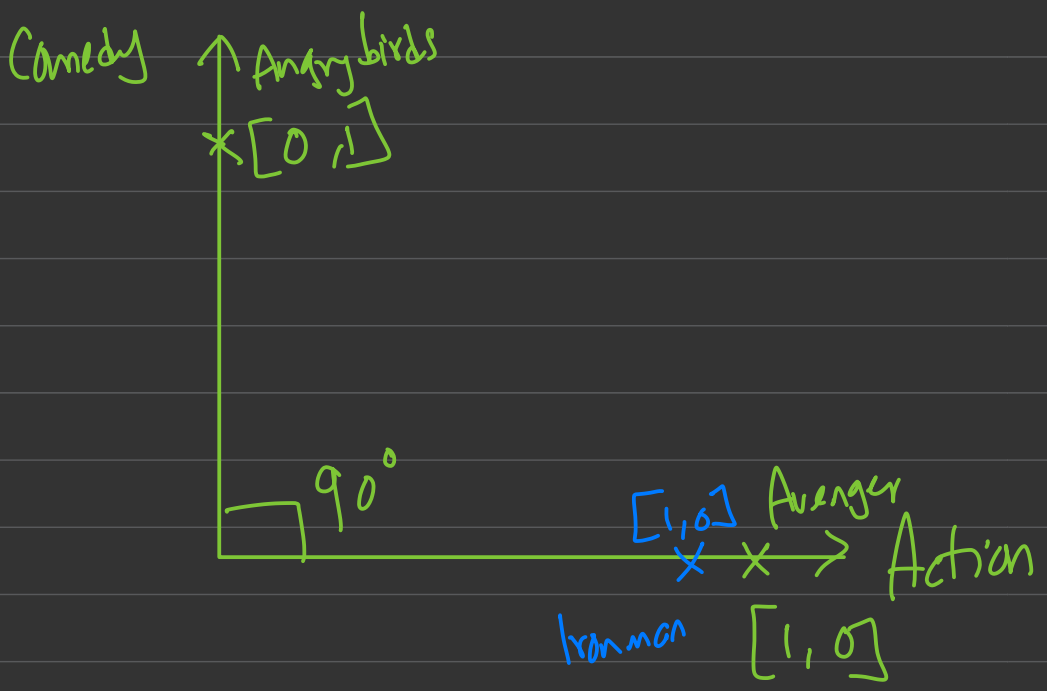
Desc. ord.

	play	love	cricket	fun
1	1	1	0	0
2	1	1	1	0
3	1	0	0	1
Σ	3	2	1	1

S1: I love playing
S2: We love to play cricket
S3: Playing is fun

→ CVect →

S1: [1 1 0 0]
S2: [1 1 1 0]
S3: [1 0 0 1]



$cos\_sim = \cos\theta$

$= \cos 90$

$= 0$

$\cos 0 = 1$

TF - IDF: Term Frequency - Inverse Document Frequency

$$TF = \frac{\#\text{ of rep of words in sentence}}{\#\text{ of words in sentence}}$$

$$IDF = \log\left(\frac{\#\text{ of sentences}}{\#\text{ of sentences containing words}}\right)$$

S1: good boy
S2: good girl
S3: boy girl good

words	Freq
good	3
boy	2
girl	2

TF	S1	S2	S3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

$*$

IDF	words	IDF
	good	$\log(3/3)=0$
	boy	$\log(3/2)$
	girl	$\log(3/2)$

	f_1 good	f_2 boy	f_3 girl
S1	0	$\frac{1}{2}\log\left(\frac{3}{2}\right)$	0
S2	0	0	$\frac{1}{2}\log\left(\frac{3}{2}\right)$
S3	0	$\frac{1}{3}\log\left(\frac{3}{2}\right)$	$\frac{1}{3}\log\left(\frac{3}{2}\right)$