How to process text data?

Input: Text data

Output: ? Search,

How to convert text into numbers so that processing is efficient, useful.

- Histogram of words.
- ...
- ...
- ...

Methods to convert text data into vectors for the purpose of ML

```
Document 1: "Machine learning is fascinating."
Document 2: "I enjoy learning about machine learning techniques."
Document 3: "Deep learning models are a subset of machine learning."
```

Method 1
- Drop the common words (stop words); Convert to lower case
- Create a dictionary.
- Express the document using the dictionary.

DICTIONARY: {machine, learning, fascinating, enjoy, techniques, deep, models, subset}

Vector embeddings of the documents
d1: {machine, learning, fascinating} = {1,1,1,0,0,0,0,0}
d2: {enjoy, learning, machine, techniques} = {machine, learning, enjoy, techniques} = {1,1,0,1,1,0,0,0}
d3: {deep, learning, models, subset, machine, learning} = {machine, learning, deep, models, subset} = {1,1,0.0.0.1.1.1}

Note:
- All documents are expressed in terms of the vocabulary
- All documents expressed in terms of vector of the same length

Applications:
- Distance between the documents can now be found out by using various measures of distance. Eg. Euclidean distance
- Similarity / difference between the documents can be established by using Cosine Similarity

Example: Distance:

$$d(d_1,d_2) = \sqrt{(1-1)^2+(1-1)^2+(1-0)^2+(0-1)^2+(0-1)^2+(0)^2+(0)^2+(0)^2}$$
$$= \sqrt{0+0+1+1+1+0+0+0} = \sqrt{3}$$
$$= 1.732$$

$$d(d_2,d_3) = \sqrt{0+0+0+1+1+1+1+1} = \sqrt{5}$$
$$= 2.236$$

$$d(d_1,d_3) = \sqrt{0+0+1+0+0+1+1+1} = \sqrt{4}$$
$$= 2$$

Limitations: - Frequency of words in the document not captured in this scheme → solution → TF-IDF.

Word vector based on frequency

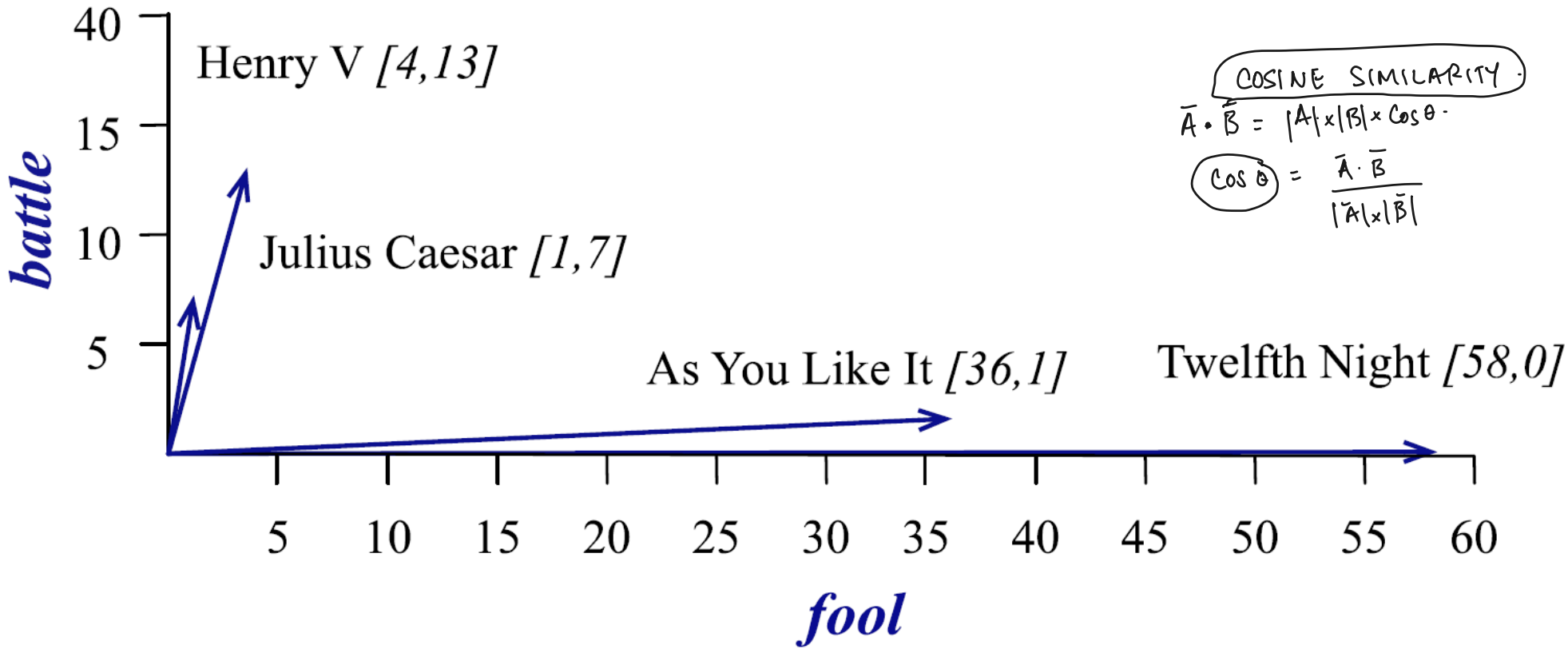
Each document is represented by a vector of words

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Vectors are similar for the two comedies

But comedies are different than the other two

Comedies have more *fools* and *wit* and fewer *battles*.



COSINE SIMILARITY

$\bar{A} \cdot \bar{B} = |A| \times |B| \times \cos\theta$

$\cos\theta = \frac{\bar{A} \cdot \bar{B}}{|\bar{A}| \times |\bar{B}|}$

Term Frequency - Inverse Document Frequency (TF-IDF)

documents

↓

Remove stock words (across the docs)

↓

Create the vocabulary (across the docs)

↓

TF-IDF

$$TF = \frac{\text{number of times this word occurred}}{\text{number of words in document}}$$

$$IDF = \log \frac{\text{number of documents}}{\text{number of documents where this word occurred}}$$

TFIDF = TF*IDF+1

```
Document 1: "Machine learning is fascinating."
Document 2: "I enjoy learning about machine learning techniques."
Document 3: "Deep learning models are a subset of machine learning."
```

Vocabulary: {"machine", "learning", "fascinating", "enjoy", "techniques", "deep", "models", "subset"}

d1: {machine, learning, fascinating}
d2: {enjoy, learning, machine, learning, techniques}
d3: {deep, learning, models, subset, machine, learning}

TF Calculations
d1: {machine: 1/3, learning: 1/3, fascinating: 1/3}
d2: {enjoy: 1/5, learning: 2/5, machine: 1/5, techniques: 1/5}
d3: {deep: 1/6; learning: 2/6; models: 1/6; subset: 1/6; machine: 1/6; learning: 1/6}

IDF Calculations for the vocabulary terms
machine: log(3/3); learning: log(3/3); fascinating: log(3/1); enjoy: log(3/1); techniques: log(3/1);
deep: log(3/1); models: log(3/1); subset: (3/1)
= {0, 0, 1.099, 1.099, 1.099, 1.099, 1.099, 1.099, 1.099}

TFIDF Calculations and vector encodings for the documents
d1: {machine: 1/3*0+1; learning: 1/3*0+1; fascinating: 1/3*1.099+1) = {1,1,1.3663,1,1,1,1,1)

d2: {machine: 1/5*0+1; learning: 2/5*0+1; fascinating:0/5*1.099+1; enjoy: 1/5*1.099+1; techniques: 1/5*1.099+1; deep: 0+1; models : 0+1; subset: 0+1} = {1, 1, 1, 1.2198, 1.2198, 1, 1, 1)

d3: {machine: 1/6*0+1; learning: 2/6*0+1; fascinating:0/6*1.099+1; enjoy: 0/6*1.099+1; techniques: 0/6*1.099+1; deep: 1/6*1.099+1; models : 1/6*1.099+1; subset: 1/6*+1.099}
= {1, 1, 1, 1,1,1.1831, 1.1831,1.1831}

TF-IDF based vector embeddings of the documents:
d1 : {1,1,1.3663,1,1,1,1,1)
d2 : {1, 1, 1, 1.2198, 1.2198, 1, 1, 1)
d3 : {1, 1, 1, 1,1,1.1831, 1.1831,1.1831}

COSINE SIMILARITY between documents
$d1 \& d2 = \frac{d1 \cdot d2}{|d1| \times |d2|}$

Applications: Distance between documents
d(d1d2) = 0.480
d(d2d3) = 0.444
d(d1d3) = 0.484

Cosine similarity:
C(d1d2) : 0.987
C(d2d3) : 0.989
C(d1d3) : 0.987

DISADVANTAGES : Contexts of the words are not considered; only their frequencies are considered.

Solution → Word2Vec.