JAMES COOK UNIVERSITY
AUSTRALIA

Week 2
MA5851 – Data Science Master Class 1
[Natural Language Processing]
[Extracting and quantifying informative text]

Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au

online.jcu.edu.au

Cairns
Singapore
Townsville

Announcements

- Collaboration session on A2
  - Tuesday, 23 March 2021, 7pm AEDT
- Weekly forum
- Quiz 1 due 22/03/2021 12:59 am

Agenda

- SLP Week 1
- Week 2 Topics
- Week 2 SLPs

SLP Week 1

- MA5851 Week 1 SLP 2
- MA5851 Week 1 SLP 3

44 submissions; total students 82

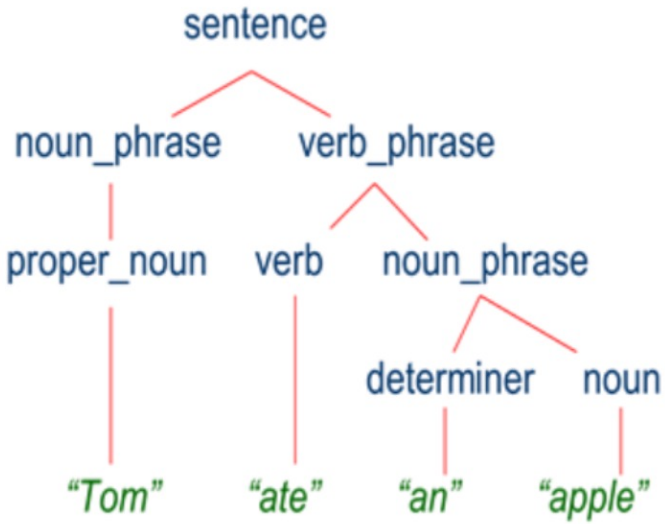
Week 2 – Extracting and Quantifying Informative Text

JAMES COOK UNIVERSITY AUSTRALIA

- Language Parsing
- Document metrics
- Zipf’s law
- Concept linking
- Quantifying text (Vectorisation, TF-IDF)
- Metadata and Search Engine
- Semantic web

Language Parsing

1. Token extraction (or 'tokenisation')
2. Term extraction (token + language = term)
3. Concept extraction (noun, noun phrases)
4. Entity extraction (associate nouns with entities)
5. Atomic fact extraction (associate nouns with verbs, that is, subject = action)
6. Complex fact extraction (e.g. natural language understanding used in AI)



"i am so disappointed that i am paying thousands of dollars for a course that has been prepared in a way where it will be largely useless in my professional career."

Parsing output:
['(NX i/NN NX) (VX am/VBP so/RB disappointed/VBN VX) that/IN (NX I/PRP NX) (VX am/VBP paying/VBG VX) (NX thousands/NNS NX) of/IN (NX dollars/NNS NX) for/IN (NX a/DT course/NN NX) (NX that/WDT NX) (VX has/VBZ been/VBN prepared/VBN VX) in/IN (NX a/DT way/NN NX) where/WRB (NX it/PRP NX) (VX will/MD be/VB largely/RB VX) (AX useless/JJ AX) in/IN (NX my/PRP$ professional/JJ career/NN NX) ./.']

Document metrics


```
{
    "content_metrics": {
        "number_of_sentences": 120,
        "number_of_paragraphs":19,
        "number_of_characters": 5447,
        "number_of_words": 2379,
        "number_of_characters_without_whitespaces": 5170,
        "number_of_noun_phrases":520,
        "number_of_verb_phrases":618,
        "list_of_np":[.......],
        "list_of_vp":[.......],
        "list_of_freq_words":[.......],
......
    }
}
```


Metric	Train	Validation	Test
No. of documents	238	10	10
No. of entries	8387	485	376
No. of phrases	21497	1329	973
No. of tokens	34718	2169	1571
Mean entries per document	35.2	48.5	37.6
Mean phrases per document	90.3	132.9	97.3
Mean tokens per document	145.9	216.9	157.1
Mean phrases per entry	2.6	2.7	2.6
Mean tokens per entry	4.1	4.5	4.2
Mean tokens per phrase	1.6	1.6	1.6
Vocabulary of target tokens	2267	442	4372
Out of vocabulary tokens	N/A	48	52

- Pattern matching: pattern of words, sentences, paragraph, punctuation, structural items etc. → document similarity
  - Plagiarism checker, stylometry, forensic legal work
- Content extraction: NP, VP, entities → summarisation

Zipf's Law

If r be the rank of word, Prob(r) be the probability of a word at rank r,
Prob(r) = freq(r) / N
freq(r) = the number of times the word at rank r appears in the collection
N = total number of words in the collection (not number of unique words).

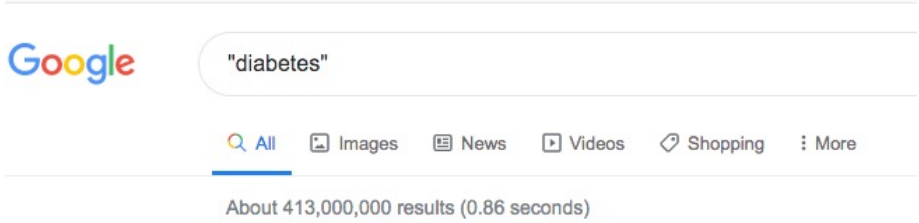
Zipf's law states that: r * Prob(r) = A; In most cases, A= 0.1
Prob(r) = freq(r) / N we can rewrite Zipf's law as
r * freq(r) = A * N
Freq(r) = (A * N)/r
the frequency of any word is inversely proportional to its rank in the frequency table

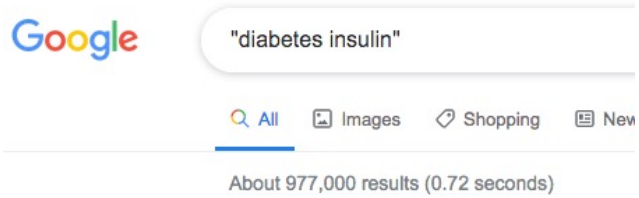
the frequency of the nth most common word is about $1/n$ times the frequency of the most common word

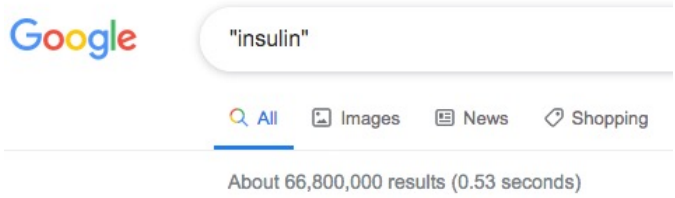
The same relationship occurs in many other rankings, unrelated to language, such as the population ranks of cities in various countries, corporation sizes, income rankings, etc.

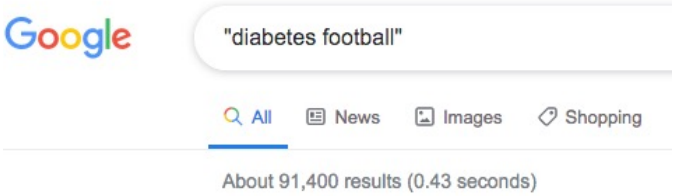
http://norvig.com/mayzner.html

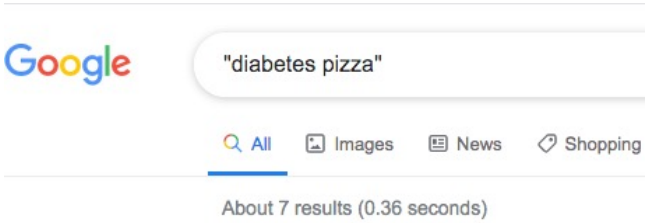
Concept linking (co-occurance)











Textual Similarity Metrics

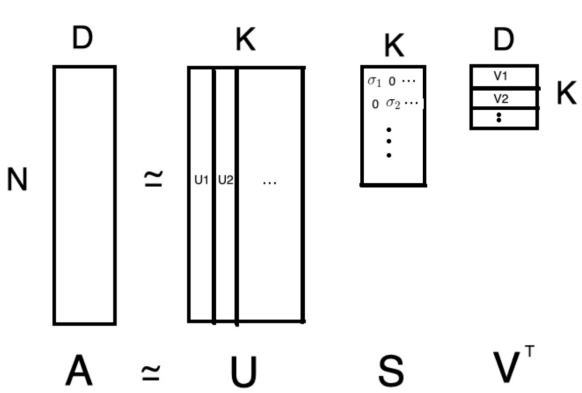
JAMES COOK UNIVERSITY AUSTRALIA

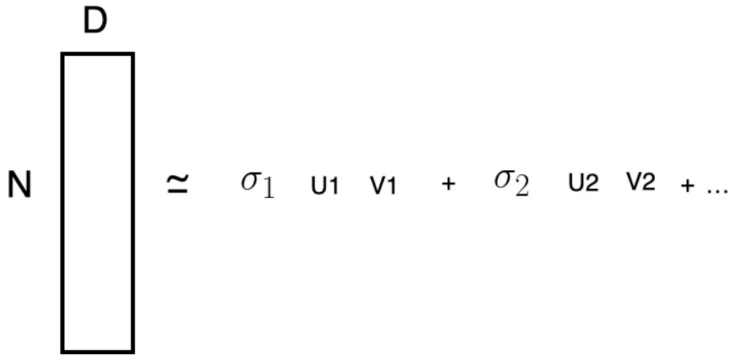
- Measuring similarity of two texts is a well-studied problem.
- Standard metrics are based on a "bag of words" model of a document that ignores word order and syntactic structure.
- May involve removing common "stop words" and stemming to reduce words to their root form.
- Vector-space model from Information Retrieval (IR) is the standard approach.
- Other metrics (e.g., Levenshtein-distance) are also used.
- Levenshtein distance: text similarity measure that compares two words and returns a numeric value representing the distance between them.
- Word Embedding: neural network based; learnt representation word (of a given corpus)

Quantifying Text

Quantifying Steps:

1. Text Corpus as input: pre-processing like tokenisation, stemming, synonyms, filtering etc.
2. Form a weighted matrix: document by term matrix
3. Apply Singular Value Decomposition (SVD): reduce sparsity
4. Apply Clustering: topic generation, summarisation, information extraction from lower-dimensional vector space





N: Number of rows, D: Number of dimensions;
SVD reduces dimensionality by selecting only the *t* largest singular values, and only keeping the first *t* columns of *U* and *V*

https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm

Document Collection

A collection of *n* documents can be represented in the vector space model by a term-document matrix.

An entry in the matrix corresponds to the "weight" of a term in the document; zero means the term has no significance in the document or it simply doesn't exist in the document.

$$
\begin{array}{c|cccc}
 & T_1 & T_2 & \ldots & T_t \\
D_1 & w_{11} & w_{21} & \ldots & w_{t1} \\
D_2 & w_{12} & w_{22} & \ldots & w_{t2} \\
\vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots \\
D_n & w_{1n} & w_{2n} & \ldots & w_{tn}
\end{array}
$$

The Vector-Space Model

JAMES COOK UNIVERSITY AUSTRALIA

Assume t distinct terms remain after preprocessing; call them index terms or the vocabulary.

These "orthogonal" terms form a vector space.

Dimension = t = |vocabulary|

Each term, i, in a document or query, j, is given a real-valued weight, w_{ij}.

Both documents and queries are expressed as t-dimensional vectors:

$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$

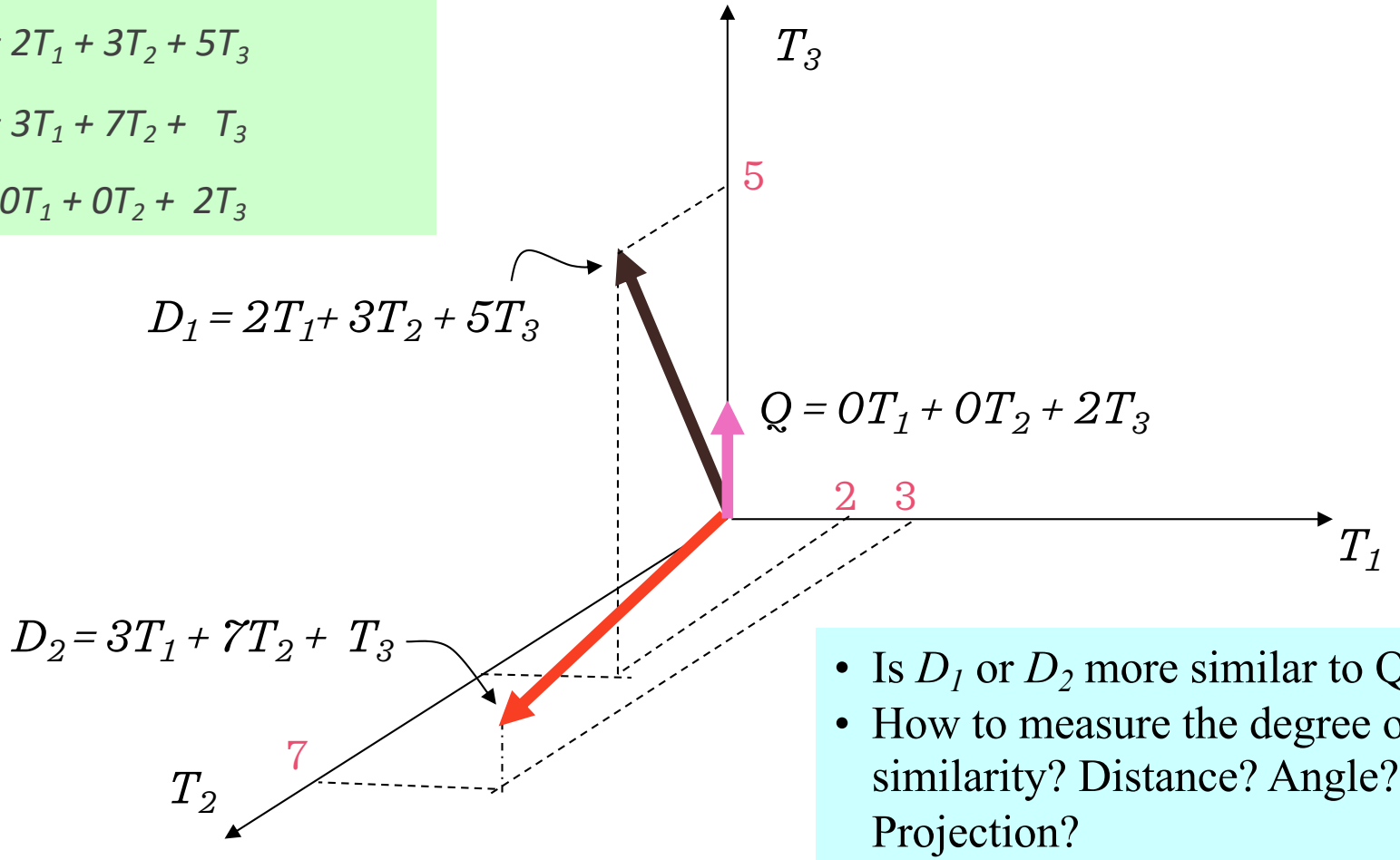
Graphic Representation

Example:

$D_1 = 2T_1 + 3T_2 + 5T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 2T_3$



- Is D_1 or D_2 more similar to Q?
- How to measure the degree of similarity? Distance? Angle? Projection?

Term Weights: Term Frequency

JAMES COOK UNIVERSITY AUSTRALIA

More frequent terms in a document are more important, i.e. more indicative of the topic.

f_{ij} = frequency of term *i* in document *j*

May want to normalize *term frequency* (*tf*) by dividing by the frequency of the most common term in the document:

$tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

Term Weights: Inverse Document Frequency

Terms that appear in many *different* documents are *less* indicative of overall topic.

df_i = document frequency of term i

= number of documents containing term i

idf_i = inverse document frequency of term i,

= $\log_2 (N/ df_i)$

(N: total number of documents)

An indication of a term's *discrimination* power.

Log used to dampen the effect relative to *tf*.

TF-IDF Weighting

JAMES COOK UNIVERSITY AUSTRALIA

A typical combined term importance indicator is *tf-idf weighting*:

$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$$

A term occurring frequently in the document but rarely in the rest of the collection is given high weight.

Many other ways of determining term weights have been proposed.

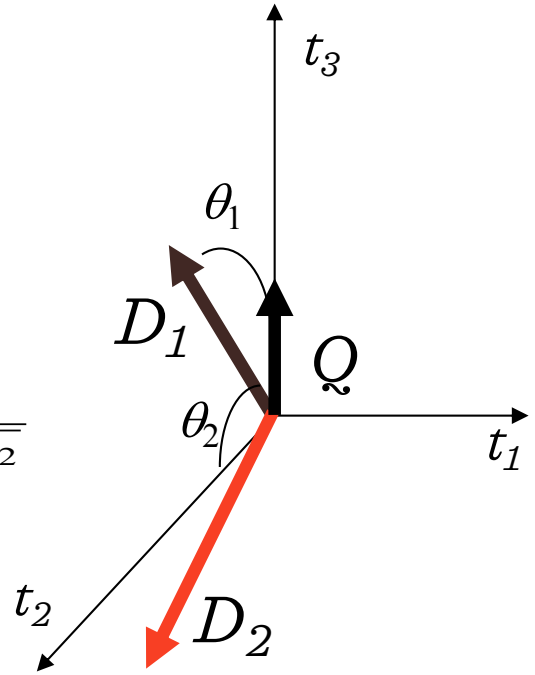
Experimentally, *tf-idf* has been found to work well.

Cosine Similarity Measure

Cosine similarity measures the cosine of the angle between two vectors.

Inner product normalized by the vector lengths.

$$\text{CosSim}(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\vec{d}_j \cdot \vec{q}}{\left|\vec{d}_j\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$



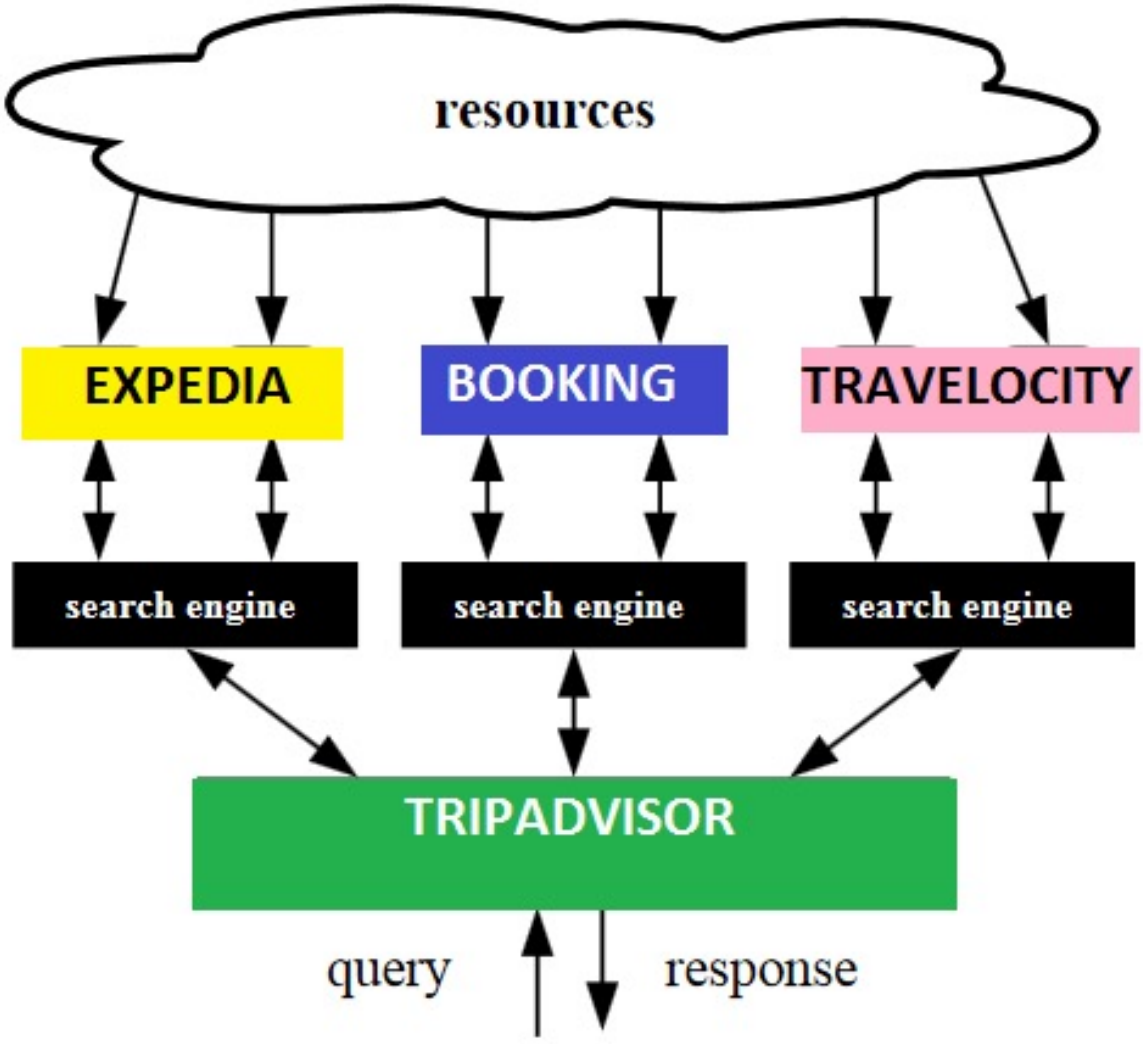
$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1 , Q) = 10 / $\sqrt{(4+9+25)(0+0+4)}$ = 0.81

$D_2 = 3T_1 + 7T_2 + 1T_3$ CosSim(D_2 , Q) = 2 / $\sqrt{(9+49+1)(0+0+4)}$ = 0.13

$Q = 0T_1 + 0T_2 + 2T_3$

D_1 is 6 times better than D_2 using cosine similarity but only 5 times better using inner product.

Metadata and Search



Semantic Web

Web of linked data.
It enables people to create data stores on the Web, build vocabularies, and write rules for handling data.
Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS

RDF: Resource Description Framework (https://www.w3.org/RDF/)
SPARQL: Query language to query RDF data
OWL: Web Ontology Language (https://www.w3.org/OWL/)
SKOS: Simple Knowledge Organization System

HL7: For Health Data, Fast Health Interop Resources (FHIR)

https://search.carrot2.org