Introduction to text mining

Preprocessing & Vector Space Model

Ayoub Bagheri

Outline

- Text mining
- Pre-processing text data
- Vector space model
  - Bag-of-words
  - Word embedding (afternoon)

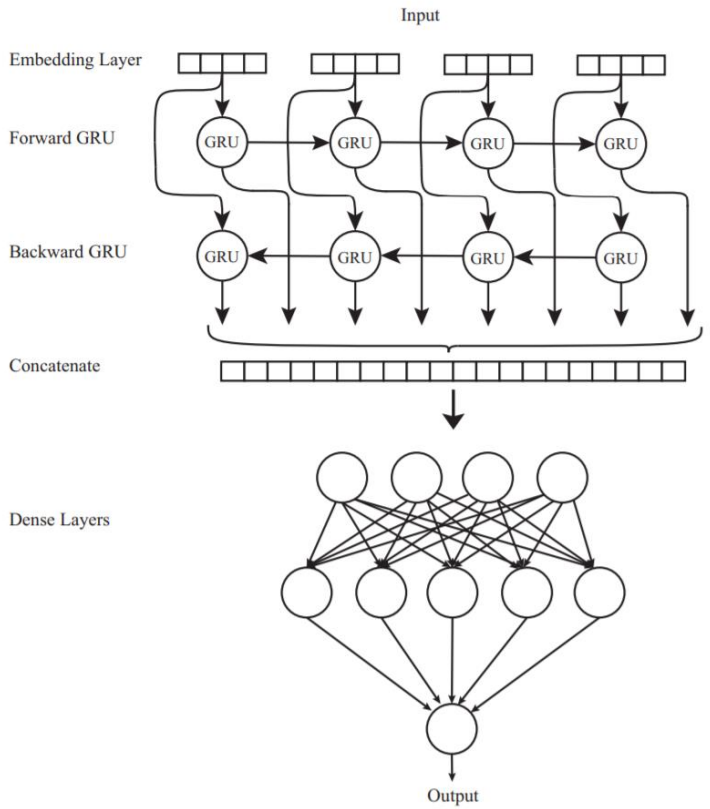
Introduction

Did a poet with donkey ears write the oldest anthem in the world?

https://dh2017.adho.org/abstracts/079/079.pdf

Automatic detection of ICD10 codes in cardiology discharge letters

https://www.nature.com/articles/s41746-021-00404-9

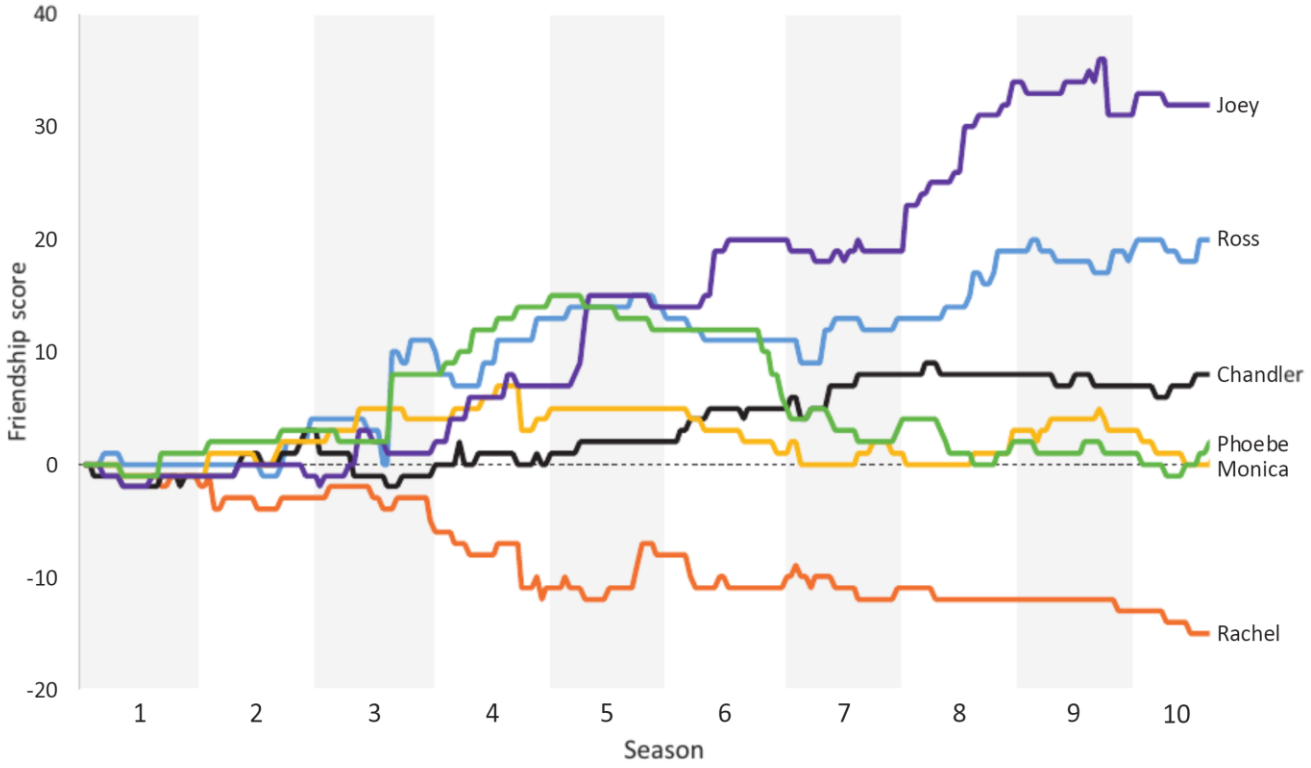


Box 1: An example of a Dutch discharge letter from the dataset

Bovengenoemde patiënt was opgenomen op <DATUM-1> op de <PERSOON-1> voor het specialisme Cardiologie.
Reden van opname *STEMI inferior*
Cardiale voorgeschiedenis*. Blanco*
Cardiovasculaire risicofactoren*: Roken(-) Diabetes(-) Hypertensie(?) Hypercholesterolemie (?)*
Anamnese*. Om 18.30 pijn op de borst met uitstraling naar de linkerarm, zweten, misselijk. Ambulance gebeld en bij aansluiten monitor beeld van acuut onderwandinfarct.*
AMBU overdracht. 500 mg aspegic iv, ticagrelor 180 mg oraal, heparine, zofran eenmalig, 3× NTG spray. HD stabiel gebleven.Medicatie bij presentatie.Geen.
Lichamelijk onderzoek*. Grauw, vegetatief, Halsvenen niet gestuwd. Cor s1 s2 geen souffles.Pulm schoon. Extr warm en slank.*
Aanvullend onderzoek*. AMBU ECG: Sinusritme, STEMI inferior III)II C/vermoedelijk RCA.*
Coronair angiografie. (…). Conclusie angio: 1-vatslijden..PCI
Conclusie en beleid
Bovengenoemde <LEEFTIJD-1> jarige man, blanco cardiale voorgeschiedenis, werd gepresenteerd vanwege een STEMI inferior waarvoor een spoed PCI werd verricht van de mid-RCA. Er bestaan geen relevante nevenletsels. Hij kon na de procedure worden overgeplaatst naar de CCU van het <INSTELLING-2>…Dank voor de snelle overname…Medicatie bij overplaatsing. Acetylsalicylzuur dispertablet 80 mg; oraal; 1× per dag 80 milligram; <DATUM-1>. Ticagrelor tablet 90 mg; oraal; 2× per dag 90 milligram; <DATUM-1>. Metoprolol tablet 50 mg; oraal; 2× per dag 25 milligram; <DATUM-1> .Atorvastatine tablet 40 mg (als ca-zout-3-water); oraal; 1× per dag 40 milligram; <DATUM-1>
Samenvatting
Hoofddiagnose: STEMI inferior wv PCI RCA. Geen nevenletsels. Nevendiagnoses: geen.
Complicaties: geen Ontslag naar: CCU <INSTELLING-2>.

Who was the best Friend?

https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1740-9713.01574



Text mining

- “The discovery by computer of **new, previously unknown information**, by automatically extracting information from different written resources”, Hearst (1999).

- Text mining is about looking for **patterns in text**, in a similar way that data mining can be loosely described as looking for patterns in data.

- Text mining describes a set of linguistic, statistical, and machine learning techniques that model and **structure the information content of textual sources**. (Wikipedia)

Why text mining?

- **Text data is everywhere**, websites (e.g., news), social media (e.g., twitter), databases (e.g., doctors' notes), digital scans of printed materials, …
- A lot of world's data is in **unstructured text format**
- Applications in industry: search, machine translation, sentiment analysis, question answering, …
- Applications in science: cognitive modeling, understanding bias in language, automated systematic literature reviews, …

Language is hard!

- Different things can mean more or less the same (“data science” vs. “statistics”)
- Context dependency (“You have very nice shoes”);
- Same words with different meanings (“to sanction”);
- Lexical ambiguity (“we saw her duck”)
- Irony, sarcasm (“You should swallow disinfectant”?)
- Figurative language (“He has a heart of stone”)
- Negation (“not good” vs. “good”), spelling variations, jargon, abbreviations
- All the above is different over languages, 99% of work is on English!

Pre-processing Text Data

Text preprocessing

- is an approach for cleaning and noise removal of text data.
- brings your text into a form that is analyzable for your task.
- transforms text into a more digestible form so that machine learning algorithms can perform better.

Why?

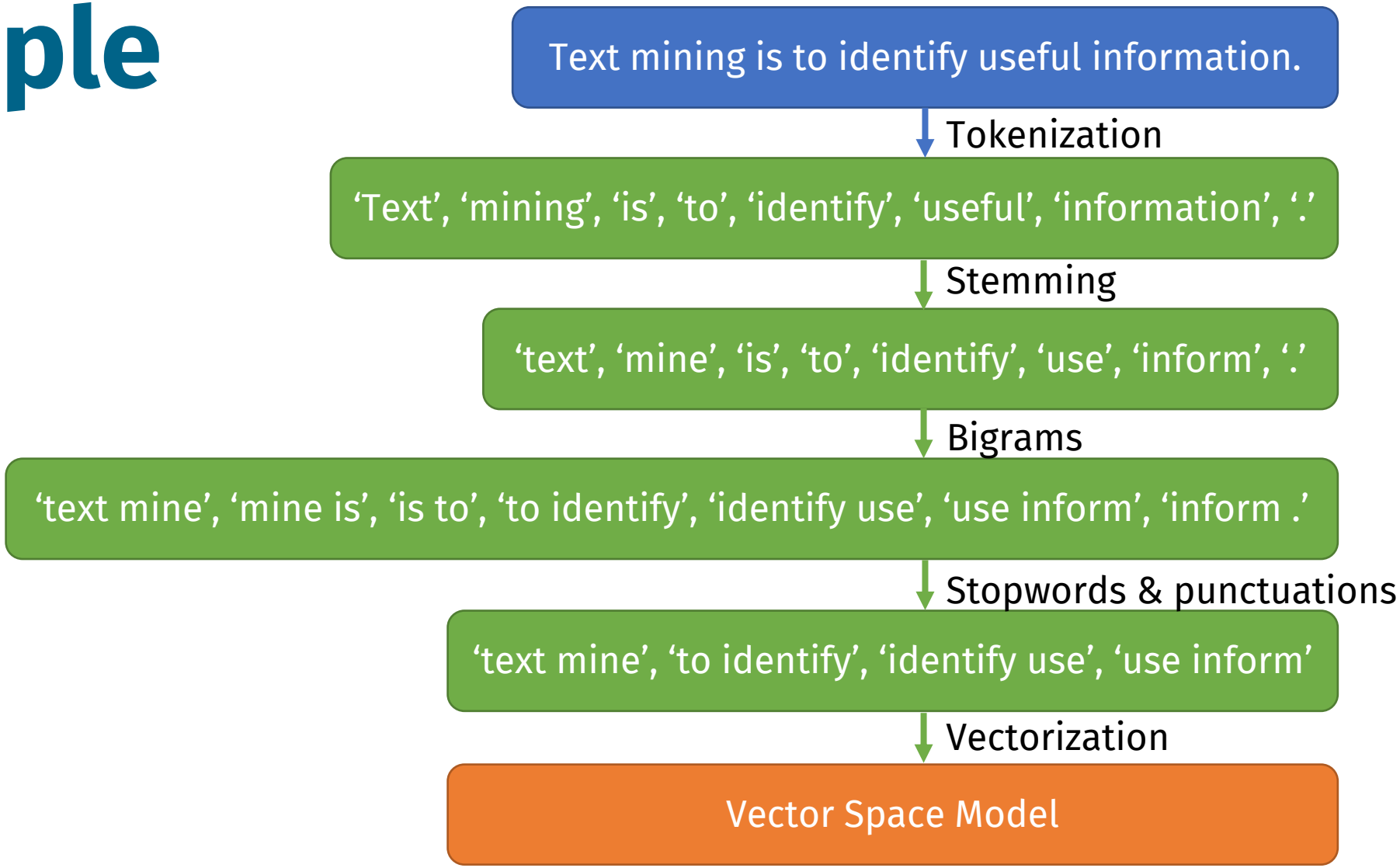
- Dimensionality
- Memory allocation
- Performance

Typical steps

- Tokenization (“text”, “ming”, “is”, “the”, “best” , “!”)
- Stemming (“running”→“run”) or Lemmatization (“were”→“is”)
- Lowercasing (“And”→“and”)
- Stopword removal (“text ming is best!”)
- Punctuation removal (“text ming is the best”)
- Number removal (“infomda 3”→“infomda”)
- Spell correction (“ming”→“mining”)

Not all of these are appropriate at all times!

Example


Text mining is to identify useful information.
Tokenization
'Text', 'mining', 'is', 'to', 'identify', 'useful', 'information', '.'
Stemming
'text', 'mine', 'is', 'to', 'identify', 'use', 'inform', '.'
Bigrams
'text mine', 'mine is', 'is to', 'to identify', 'identify use', 'use inform', 'inform .'
Stopwords & punctuations
'text mine', 'to identify', 'identify use', 'use inform'
Vectorization
Vector Space Model

Vector Space Model

Basic idea

- Text is “unstructured data”
- How do we get to something structured that we can compute with?
- **Text must be represented somehow**
- Represent the text as something that makes sense to a computer

Vector space model

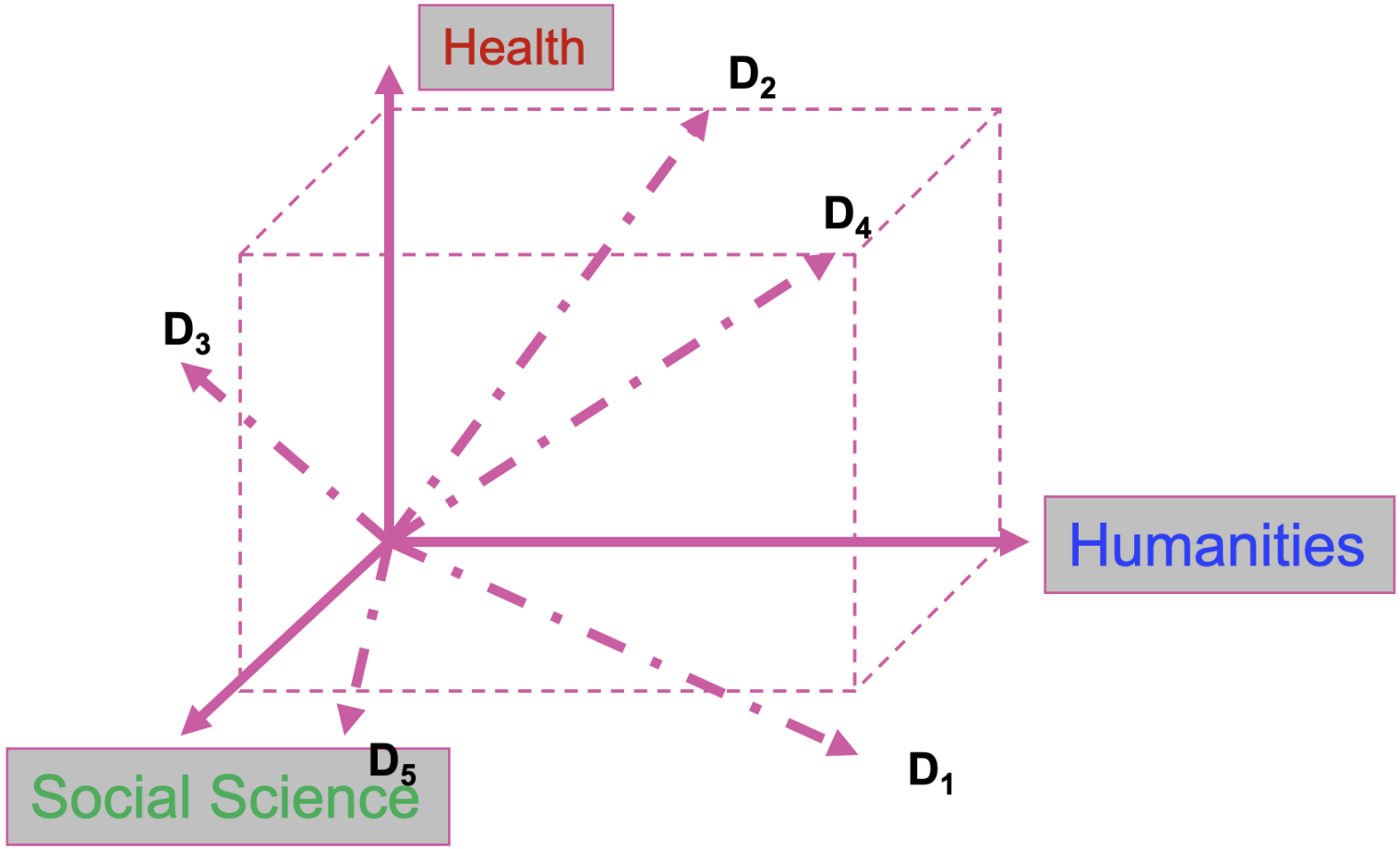
- A vector space is a collection of vectors
- Represent documents by concept vectors
  - Each concept defines one dimension
  - k concepts define a high-dimensional space
  - Element of vector corresponds to concept weight

Vector space model

- Terms / words are generic features that can be extracted from text
- Typically, terms are single words, keywords, n-grams, or phrases
- Documents are represented as vectors of terms
- Each dimension (concept) corresponds to a separate term

$$d = (w_1, \ldots, w_n)$$

An illustration

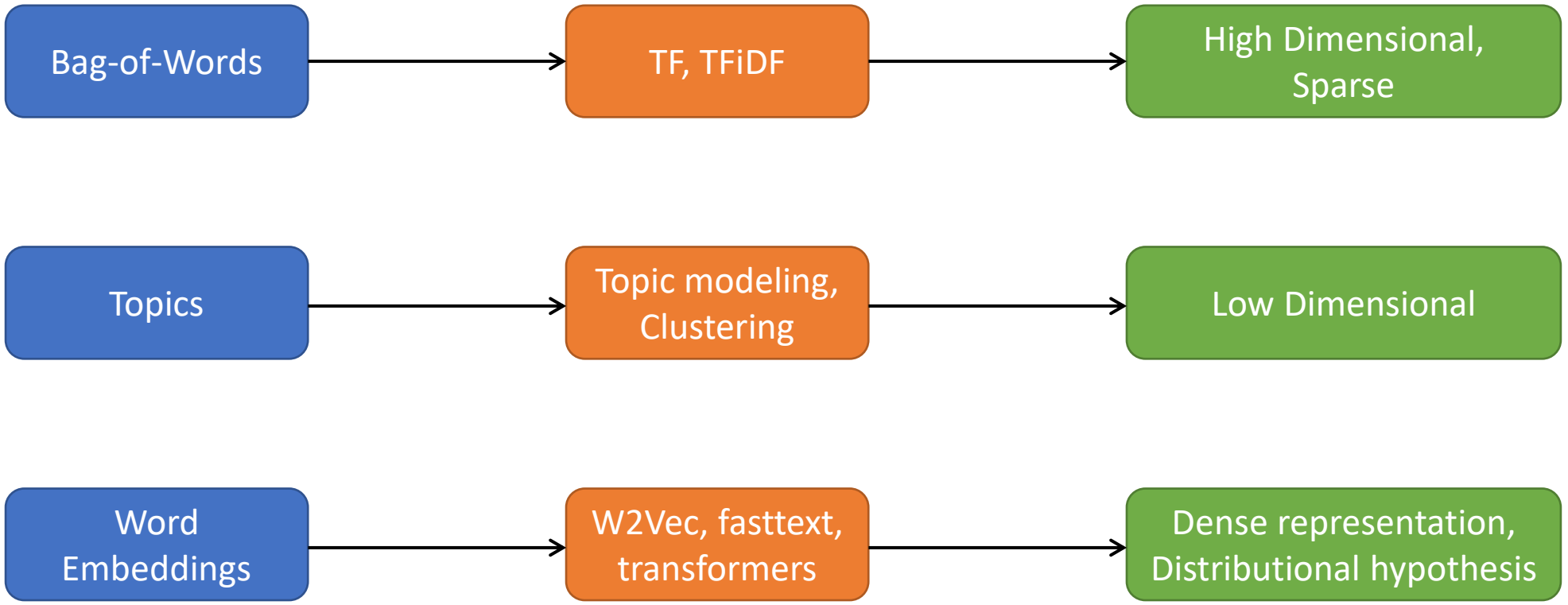


Vectorization

- The process of converting text into numbers is called **Vectorization**

- Distance between the vectors in this concept space
  - Relationship among documents

VSM representations



Bag-of-Words

- **Terms** are words (more generally we can use n-grams)
- **Weights** are number of occurrences of the terms in the document
  - Binary
  - Term Frequency (TF)
  - Term Frequency inverse Document Frequency (TFiDF)

Example

Doc1: Text mining is to identify useful information.

Doc2: Useful information is mined from text.

Doc3: Apple is delicious.

Document-Term matrix (DTM):

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

TFiDF

- A term is more discriminative if it occurs a lot but only in fewer documents
- Let $n_{d,t}$ denote the number of times the t-th term appears in the d-th document.

$$TF_{d,t} = \frac{n_{d,t}}{\sum_i n_{d,i}}$$

- Let N denote the number of documents and N_t denote the number of documents containing the t-th term.

$$IDF_t = log(\frac{N}{N_t})$$

TFiDF weight:

$$w_{d,t} = TF_{d,t} \cdot IDF_t$$

Similarity metric

- Euclidean distance
  - $dist(d_i, d_j) = \sqrt{\sum_{t \in V} [tf(t, d_i) idf(t) - tf(t, d_j) idf(t)]^2}$
  - Longer documents will be penalized by the extra words
  - We care more about how these two vectors are overlapped

Text data are High Dimensional!

Terminology

- **Corpus:** is a large and structured set of texts
- **Stop words:** words which are filtered out before or after processing of natural language data (text)
- **Unstructured text:** information that either does not have a pre-defined data model or is not organized in a pre-defined manner.
- **Tokenizing:** process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens (see also lexical analysis)
- **Natural language processing:** field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages
- **Term document (or document term) matrix:** is a mathematical matrix that describes the frequency of terms that occur in a collection of documents
- **Supervised learning:** s the machine learning task of inferring a function from labeled training data
- **Unsupervised learning:** find hidden structure in unlabeled data
- **Stemming:** the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form

Practical

Preprocessing and create document-term matrices on BBC news dataset.

Questions?