SS-ZG548: ADVANCED DATA MINING

12 Text Mining

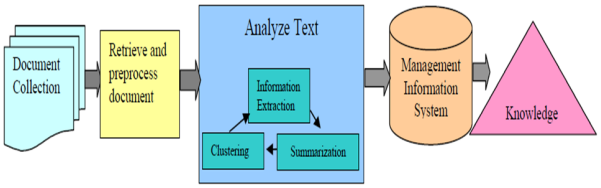
Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Oct 25, 2021 ONLINE (WILP @ BITS-Pilani July-Dec 2021)

http://ktiwari.in/adm

Text Mining

- Computer are bad to handle slang, spelling variations, contextual meaning and unstructured data
- Text is less structured
- Applications involves 1) Information Extraction, 2) Topic Tracking, 3) Summarization, 4) Categorization, 5) Clustering, 6) Concept Linkage, 7) Information Visualization, 8) Question Answering, *etc.*
- Starts with 1) Identity keywords and phrases, 2) Relationship within text



Text Representation

- Binary term-document incidence matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Document is represented as a binary vector $\in \{0,1\}^{|V|}$

- Term-document count matrices

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Document is represented as a count vector $\in N^{|V|}$

Text Feature

Dictionary: Indian , cricketer, captain , team, batsmen, run, player, centuries, century, ODI, international , records, politician , Prime Minister, Chief Minister, Member of Parliament, General Secretary, constituency

1. Sachin Ramesh Tendulkar is a former Indian international cricketer and a former captain of the Indian national team, regarded as one of the greatest batsmen of all time. He is the highest run scorer of all time in International cricket. Tendulkar took up cricket at the age of eleven, made his Test debut on 15 November 1989 against Pakistan in Karachi at the age of sixteen, and went on to represent Mumbai domestically and India internationally for close to twenty-four years. He is the only player to have scored one hundred international centuries, the first batsman to score a double century in a ODI, the holder of the record for the most number of runs in both Test and ODI, and the only player to complete more than 30,000 runs in international cricket.
2. Rahul Gandhi is an Indian politician. He hails from a long line of politicians, known as the Nehru-Gandhi family, which has occupied a prominent place in the politics of India ever since the country gained independence in 1947. His great-grandfather was Jawaharlal Nehru, the first prime minister of India and also the longest serving Prime Minister of India having served for a total of seventeen years. The son of Sonia and Rajiv Gandhi, he is the President of the Indian National Congress and serves such additional offices as the Chairperson of the Indian Youth Congress and the National Students Union of India. A member of the Indian Parliament, Gandhi represents the constituency of Amethi, Uttar Pradesh.
3. Sunil Manohar Gavaskar is a former Indian international cricketer who played from the early 1970s to late 1980s for the Mumbai cricket team and Indian national team. Widely regarded as one of the greatest Test batsmen and best opening batsmen in Test cricket history, Gavaskar set world records during his career for the most Test runs and most Test centuries scored by any batsman. He held the record of 34 Test centuries for almost two decades before it was broken by Sachin Tendulkar in December 2005. He was the first person to score centuries in both innings of a Test match three times. He was the first Test batsman to score 10,000 Test Runs in a Career and now stands at number 12 on the group of 13 players with 10,000+ Test Runs.
4. Narendra Damodardas Modi is an Indian politician serving as the 14th and current Prime Minister of India. He was the Chief Minister of Gujarat and is the Member of Parliament for Varanasi constituency. Modi helped his father sell tea as a child and later ran his own stall. He was introduced to the RSS at the age of eight, beginning a long association with the organisation. He left home after graduating from school. Modi travelled around India for two years and visited a number of religious centres. In 1971 he became a full-time worker for the RSS. During the state of emergency imposed across the country in 1975, Modi was forced to go into hiding. The RSS assigned him to the BJP in 1985, and he held several positions within the party hierarchy until 2001, rising to the rank of General Secretary.

Text Representation

- Bag of words: order of words is not important. See “Jon is lighter than Bob” and “Bob is lighter than Jon”
- Term frequency ($tf_{t,d}$): is number of times the term t occurs in document d. Relevance may not increase proportionally with term frequency.
- Log-frequency weighting: $w_{t,d} = \log(1 + tf_{t,d})$ if $tf_{t,d} > 0$ otherwise zero. Consider following matching score b/w two documents

$$score = \sum_{t \in D_1 \cap D_2} w_{t,D_1}$$

Score is zero if no term of query document D_2 is present in D_1.

Text Representation

- **Document frequency:** Frequent terms are less informative than rare terms. df_t is number of documents that contain term t.
- **Inverse document frequency:** If we have N documents then $idf_t = \log(N/df_t)$
- **Collection frequency:** How many times term t appeared in all the document.
- **tfidf weighting:** $tfidf_{t,d} = \log(1 + tf_{t,d}) \times \log(N/df_t)$

$$score = \sum_{t \in q \cap d} tfidf_{t,d}$$

This is the most used method to determine similarity.

Text Representation

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

- Generally dimensionality reduction is also required
- How document classification? use k-NN, Naive Bayes, SVM ...
- How document clustering? use k-Means, Hierarchical, Agglomerative

Thank You!

Thank you very much for your attention!

Queries ?