Sentiment Analysis (**PML** Ch. 8)
Machine Learning for Finance (FIN 570)

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2023-24 Module 3 (Spring 2024)

Overview

- Sentiment analysis (opinion mining) as a sub-field of Natural language processing (NLP)
- Classify or score the emotion or opinion in text documents for a specific purpose.
- More advanced than simple word counting (text mining). E.g.,

$$\text{Score} = \frac{\#\text{ of positive words} - \#\text{ of negative words}}{\#\text{ of positive words} + \#\text{ of negative words}},$$

  where **positive** and **negative** words are predefined for the purpose of the problem.
- A key task is to extract the numerical feature (vector) from text so that the numerical features are used as an input to ML algorithm.

Bag-of-words model

Term frequency

Create a dictionary (vocabulary) of unique tokens (e.g., words).

$$tf(t, d) = \text{the frequency of a term } t \text{ appearing in document } d.$$

- The words frequently repeated in a document is important.
- The vector size is the number of all unique words in all documents.
- The **term** can be generalized to the sequence of multiple words.
  - **1-gram model**: one word. E.g., "the", "sun", "is", "shining"
  - **2-gram model**: two words. E.g., "the sun", "sun is", "is shining"

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0

$$\texttt{tf}(t, d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}}$$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

Inverse document frequency ($idf(t, D)$)

$$idf(t) = \log \frac{n_D}{1 + df(t)} \quad \text{or} \quad \log \frac{1 + n_D}{1 + df(t)},$$

where $df(t)$ is **document frequency**, # of documents containing the term t.

- A word common across many documents is less important/relevant.
- The second definition ensures that $idf(t) = \log \frac{1+0}{1+0} = 0$ if a term t never appears in any document.

$f_{t,d}$

	blue	bright	can	see	shining	sky	sun	today
1	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1
3	0	1	0	0	0	1	1	0
4	0	1	1	1	1	0	2	0
n_t	1	3	1	1	1	2	3	1

$N = 4$

$$\texttt{idf}(t, D) = \log_{10} \frac{N}{n_t}$$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$\log_{10} \frac{4}{1} = 0.602$ $\log_{10} \frac{4}{3} = 0.125$

Term frequency-inverse document frequency (tf-idf)

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(d) \quad \text{or} \quad tf(t,d) \times \big(1 + idf(d)\big)$$

- Either $tf(t,d)$ or tf-idf(t,d) is normalized to $\|\boldsymbol{v}\| = 1$.
- tf-idf is used as input to ML algorithms.

$\texttt{tf}(t,d)$

	blue	bright	can	see	shining	sky	sun	today
1	1/2	0	0	0	0	1/2	0	0
2	0	1/3	0	0	0	0	1/3	1/3
3	0	1/3	0	0	0	1/3	1/3	0
4	0	1/6	1/6	1/6	1/6	0	1/3	0

- TF-IDF: Multiply TF and IDF scores, use to rank importance of words within documents
  - Most important word for each document is highlighted

x

$\texttt{idf}(t,D)$

	blue	bright	can	see	shining	sky	sun	today
	0.602	0.125	0.602	0.602	0.602	0.301	0.125	0.602

$\texttt{tfidf}(t,d,D) = \texttt{tf}(t,d) \cdot \texttt{idf}(t,D)$

	blue	bright	can	see	shining	sky	sun	today
1	**0.301**	0	0	0	0	0.151	0	0
2	0	0.0417	0	0	0	0	0.0417	**0.201**
3	0	0.0417	0	0	0	**0.100**	0.0417	0
4	0	0.0209	**0.100**	**0.100**	**0.100**	0	0.0417	0

Pre-processing for tokenizing

- **Stemming**: E.g., run, runs, ran $\Rightarrow$ run
- **Removing stop words**: Remove e.g. "is", "and", "had", and "like".

Other advanced methods

The **word2vec** model

- A modern version of the bag-of-words model developed in Google.
- A neural-network based unsupervised learning to learn relationship between words.
- Eventually create vectors (embedding) representing words such that related word pair has high cosine similarity.

Latent Dirichlet Allocation (DLA)

- A popular technique for **topic modeling**.
- Topic modeling group text (news article) with category labels (e.g., finance, politics, local)
- Bag-of-word matrix $\Rightarrow$ LDA $\Rightarrow$ topic words.
- Topic words must be predefined.

Application in Finance: 1. Product similarity

- Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. Journal of Political Economy, 124(5), 1423–1465. https://doi.org/10.1086/688176
- Analyze firms' business description sections from SEC 10-K filing.
- Construct word vectors (simple version of bag-of-words):
  - Create a dictionary of nouns and proper nouns used in all 10-K.
  - Omit the common words used in 20% or more firms (c.f., inverse document frequency).
  - Obtain the word vector $\boldsymbol{v}_i$ with component values being 1 (not used) or 0 (used).
- Calculate the product similarity for a firm pair (f_i, f_j) as:

$$K(f_i, f_j) = \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{|\boldsymbol{v}_i| \cdot |\boldsymbol{v}_j|} = \cos\theta_{ij}$$

- Define the text-based network industry classifications (TNIC). TNIC works better than classical industry classification (NAICS or SIC).
- Similarity data (in different levels) available in Hoberg-Phillips Data Library

Application in Finance: 2. Measuring corporate culture

- Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. The Review of Financial Studies, 34(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079
- Use **word2vec** method to analyze firm's earnings call transcripts.
- Measures 5 corporate cultural values: *innovation, integrity, quality, respect, and teamwork*.
  - Begin with the seed words of each value. E.g., **teamwork**: collaborate, collaboration, collaborative, cooperate, cooperation, cooperative, and teamwork. Calculate the average of the word vectors of the seed words.
  - Find the top 500 words (culture dictionary) with the highest similarity to the average.
  - Measure the corporate culture value as the sum of $tf\text{-}idf(d,t)$ for the words t in the each culture dictionary.

Table 2
Thirty most representative and most frequently occurring words in the culture dictionary

A. Thirty most representative words for each cultural value in the culture dictionary

Innovation	Integrity	Quality	Respect	Teamwork
Creativity	Accountability	Dedicated	Talented	Collaborate
Innovative	Ethic	Quality	Talent	Cooperation
Innovate	Integrity	Dedication	Empower	Collaboration
Innovation	Responsibility	Customer_service	Team_member	Collaborative
Creative	Transparency	Customer	Employee	Cooperative
Excellence	Accountable	Dedicate	Team	Partnership
Passion	Governance	Service_level	Leadership	Cooperate
World-class	Ethical	Mission	Leadership_team	Collaboratively
Technology	Transparent	Service_delivery	Culture	Partner
Operational_excellence	Trust	Customer_satisfaction	Teammate	Co-operation
Passionate	Responsible	Service	Organization	Coordination
Product_innovation	Oversight	Reliability	Entrepreneurial	Engage
Capability	Independence	Commitment	Skill	Jointly
Customer_experience	Objectivity	Customer_need	Executive	Coordinate
Thought_leadership	Moral	Customer_support	Empowerment	Teamwork
Expertise	Trustworthy	High-quality	Management_team	Business_partner
Agility	Fairness	Ensure	Best_brightest	Alliance
Efficient	Hold_accountable	Customer_relationship	Professionalism	Team_up
Technology_innovation	Corporate_governance	Quality_service	Staff	Technology_partner
Competency	Autonomy	Product_quality	Highly_skilled	Joint
Know-how	Core_value	Quality_product	Skill_set	Cooperatively
Cutting-edge	Assure	Capable	Technologist	Relationship
Agile	Stakeholder	Service_quality	Competent	Collaborator
Creatively	Fiduciary_responsibility	End_user	Entrepreneur	Interaction
Customer-centric	Continuity	Quality_level	Experienced	Working_relationship
Enable	Credibility	Customer_expectation	Energize	Co-operate
Value_proposition	Honesty	Service_capability	Entrepreneurial_spirit	Technology_partnership
Reinvent	Privacy	Client	High-caliber	Association
Focus	Fiduciary_duty	Customer_requirement	Manager	Dialogue
Innovation_capability	Rigor	Sla	Leadership_skill	Dialog