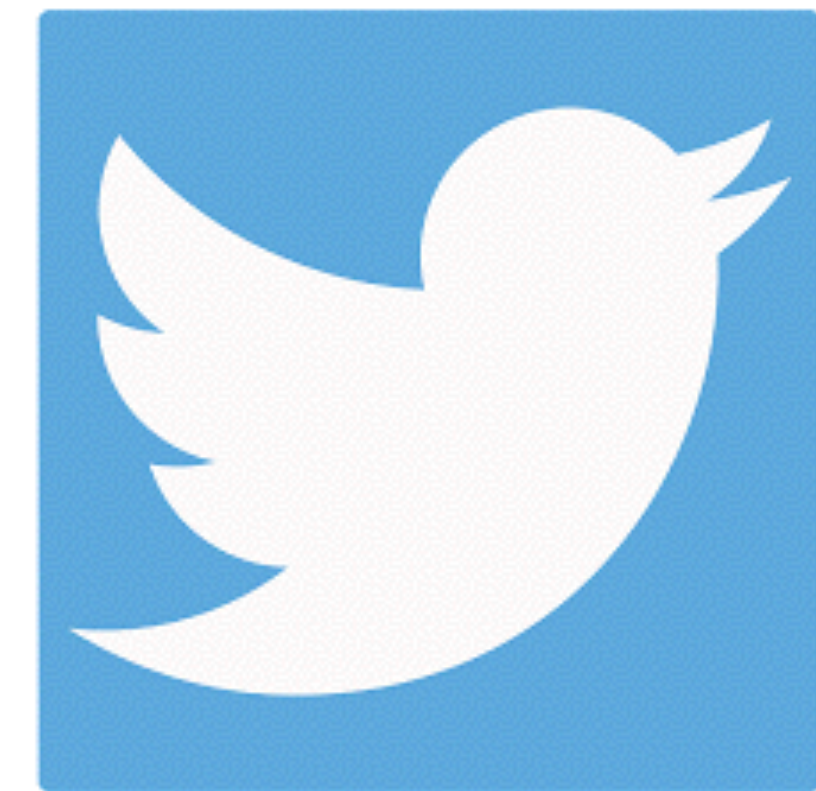
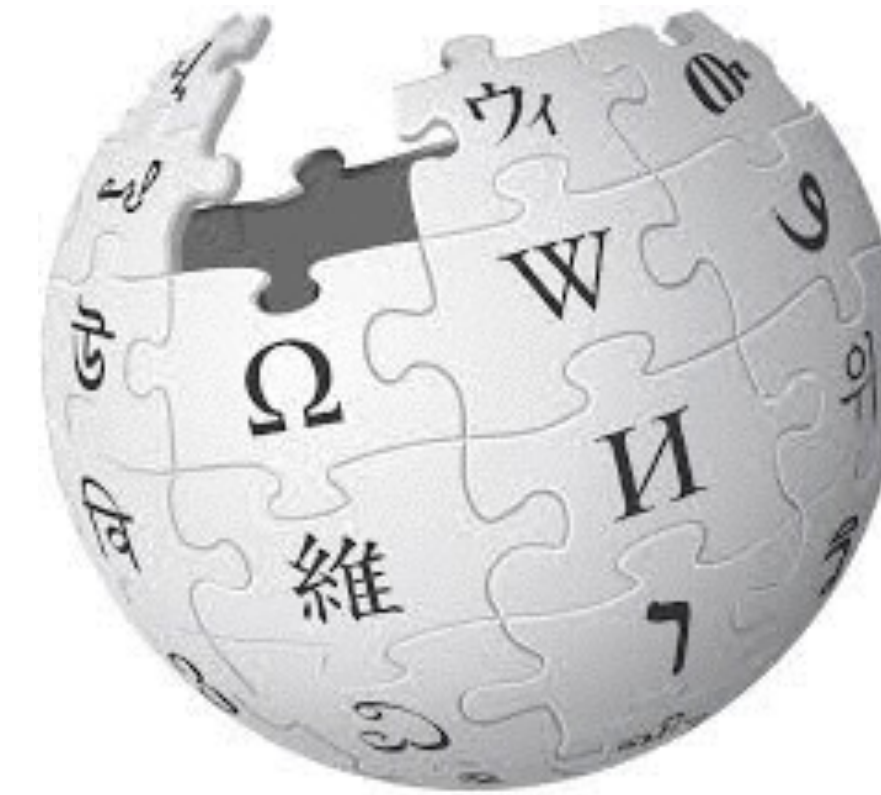


Applied Text Mining in Python

Introduction to Text Mining

Text is Everywhere!



Text data is growing fast!

- **Data continues to grow exponentially**
 - Estimated to be 2.5 Exabytes (2.5 million TB) a day
 - Grow to 40 Zettabytes (40 billion TB) by 2020 (50-times that of 2010)
- **Approximately 80% of all data is estimated to be unstructured, text-rich data**
 - >40 million articles (5 million in English) in Wikipedia
 - >4.5 billion Web pages
 - >500 million tweets a day, 200 billion a year
 - >1.5 trillion queries / searches on Google a year

Data hidden in plain sight

Social
network

Author

Description

Location

Tweet

- Topic
- Sentiment

Time

Popularity

The image shows a screenshot of the Twitter profile for the UN Spokesperson (@UN_Spokesperson). The profile picture is the United Nations logo. The header image shows two men in suits. The profile information includes the name 'UN Spokesperson', the handle '@UN_Spokesperson', a description 'Official Twitter account of the Office of the Spokesperson for United Nations Secretary-General Ban Ki-moon.', location 'New York, USA', website 'un.org/sg/spokesperso...', and 'Joined May 2010'. The stats show 14.6K tweets, 994 following, 391K followers, 49 likes, and 3 lists. The 'Follow' button is visible. Below the profile information, there are three tweets. The first tweet is about maintaining unity on the Korean Peninsula. The second tweet is about ethics being built into the ideals and objectives of the United Nations. The third tweet is about Ban on Amb. Joseph V. Reed. The engagement metrics for the second tweet (6 replies, 13 retweets, 27 likes) are circled in red. Arrows from the surrounding text boxes point to various elements: 'Social network' points to the profile picture, 'Author' points to the name, 'Description' points to the bio, 'Location' points to the location, 'Tweet' points to the tweet text, 'Topic' points to the hashtags, 'Sentiment' points to the tweet text, 'Time' points to the time relative to the tweet, and 'Popularity' points to the engagement metrics.

UN Spokesperson @UN_Spokesperson
Official Twitter account of the Office of the Spokesperson for United Nations Secretary-General Ban Ki-moon.
New York, USA
un.org/sg/spokesperso...
Joined May 2010

Tweets Tweets & replies Media

UN Spokesperson @UN_Spokesperson · 3h
Maintaining unity is crucial in tackling security challenges on Korean Peninsula & beyond: #UNSG on #DPRK sanctions bit.ly/2gVeX7z

UN Spokesperson @UN_Spokesperson · 17h
"Ethics are built right into the ideals and objectives of the United Nations" #UNSG @ NY Society for Ethical Culture bit.ly/2guVelr

UN Spokesperson @UN_Spokesperson · 23h
Ban on Amb. Joseph V. Reed: The UN family is fortunate to have had such a wonderful supporter, wonderful leader. bit.ly/2gFU1yp

So, what can be done with text?

- Parse text
- Find / Identify / Extract relevant information from text
- Classify text documents
- Search for relevant text documents
- Sentiment analysis
- Topic modeling
- ...