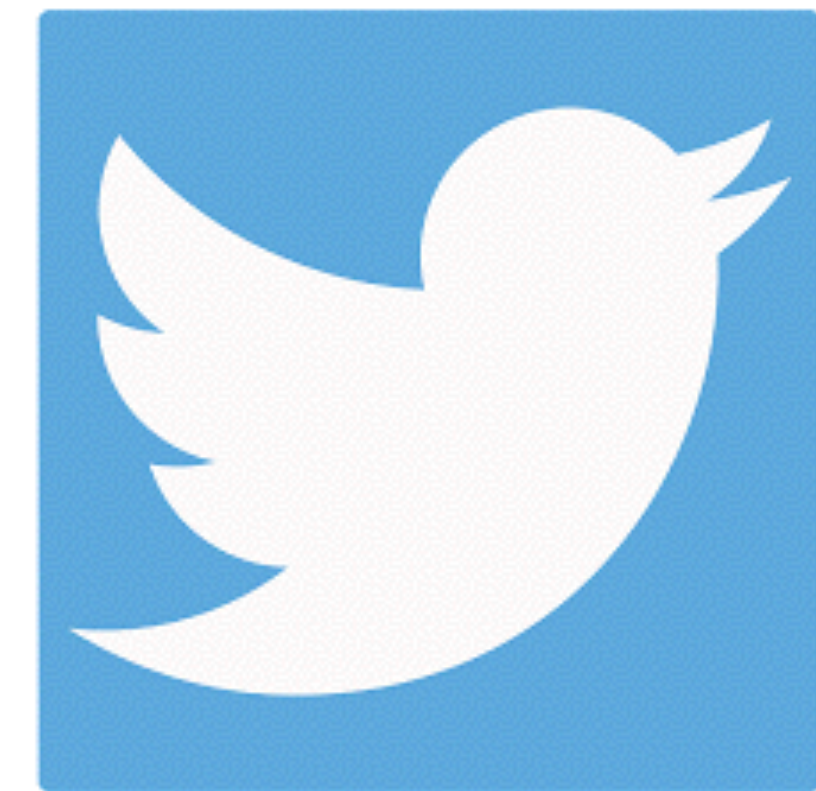
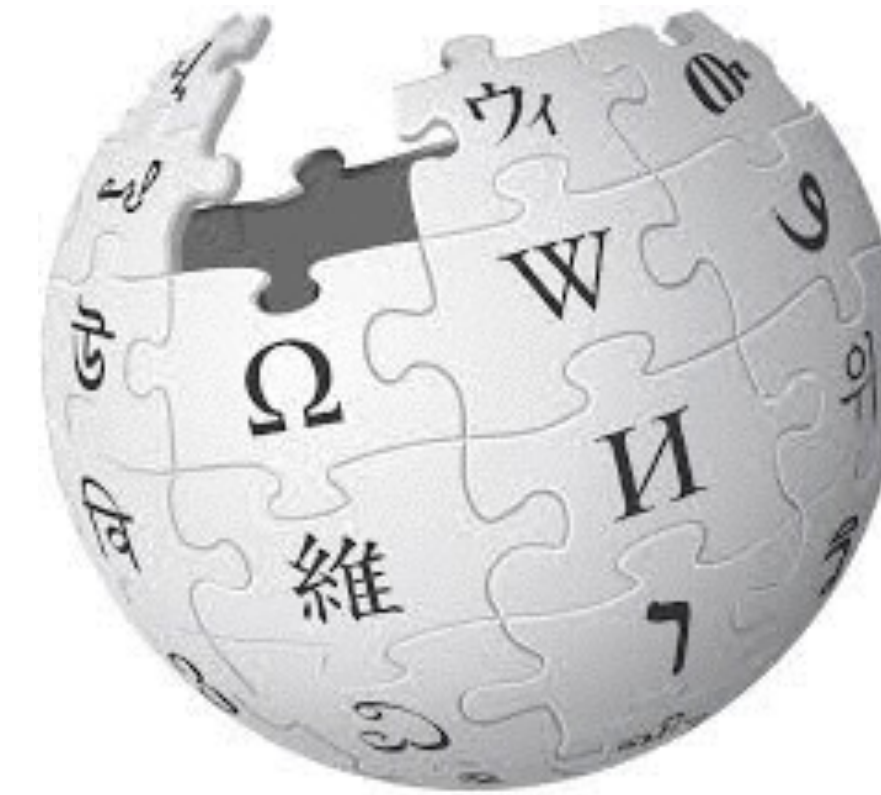


Applied Text Mining in Python

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 UN Spokesperson's Twitter profile. The profile header includes the UN logo, the name 'UN Spokesperson' with a verified badge, and the handle '@UN_Spokesperson'. The bio states: 'Official Twitter account of the Office of the Spokesperson for United Nations Secretary-General Ban Ki-moon.' The location is 'New York, USA', the website is 'un.org/sg/spokesperso...', and it was joined in 'May 2010'. There is a button to 'Tweet to UN Spokesperson' and a link to '3,008 Photos and videos'.

The 'Tweets' tab is selected, showing three tweets. The first tweet is from 3 hours ago, discussing security challenges on the Korean Peninsula. The second tweet is from 17 hours ago, discussing ethics at the NY Society for Ethical Culture. The third tweet is from 23 hours ago, discussing the ban on Ambassador Joseph V. Reed.

Annotations highlight specific data points:

- Social network:** Points to the UN logo in the profile header.
- Author:** Points to the profile name 'UN Spokesperson'.
- Description:** Points to the bio text.
- Location:** Points to the location 'New York, USA'.
- Tweet:** Points to the text of the second tweet.
- Time:** Points to the time '17h' of the second tweet.
- Popularity:** Points to the engagement metrics (6 replies, 13 retweets, 27 likes) of the second tweet.

Tweets	Following	Followers	Likes	Lists
14.6K	994	391K	49	3

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
- ...