Applied Text Mining in Python

Introduction to Text Mining

Text is Everywhere!

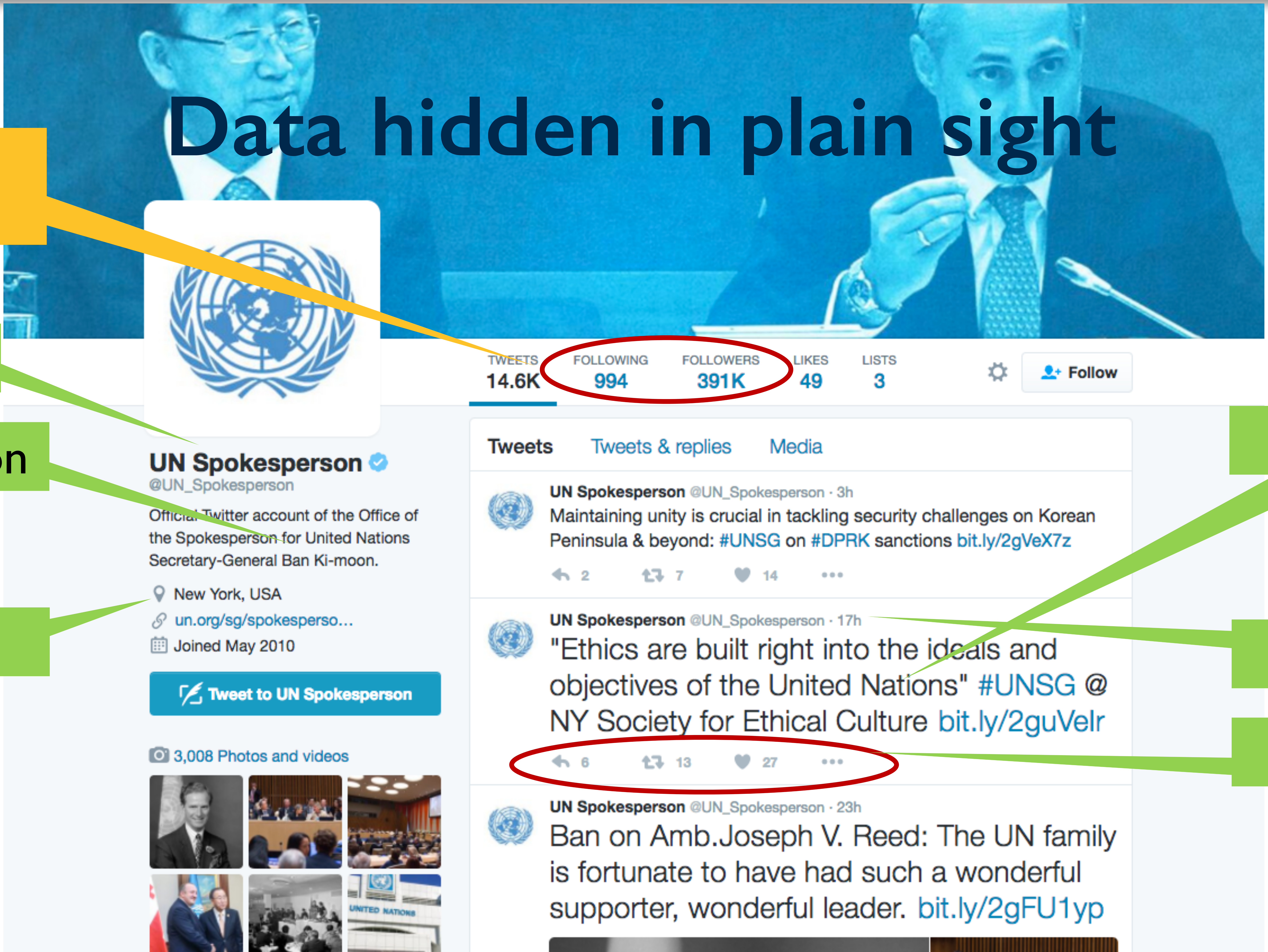
Dec 29, 2004 - The Inkscape developers have set a goal for the release of Inkscape 0.41 for mid-late January. There are some noteworthy new features that would be worth getting out to users, and with some major internal changes that will be taking place soon, a good stable release is needed prior to starting such an undertaking.

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



TWEETS 14.6K | FOLLOWING 994 | FOLLOWERS 391K | LIKES 49 | LISTS 3

Follow

Tweets | Tweets & replies | Media

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

Tweet to UN Spokesperson

3,008 Photos and videos

UN Spokesperson @UN_Spokesperson · 3h
Maintaining unity is crucial in tackling security challenges on Korean Peninsula & beyond: #UNSG on #DPRK sanctions bit.ly/2gVeX7z
2 7 14

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
6 13 27

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

Social network

Author

Description

Location

Tweet
- Topic
- Sentiment

Time

Popularity

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
- …