Text Mining - Exam subject (1)

- Dataset to be used
  - lupus.csv (506 lines), extracted from Medline database: a review of papers published between 1994 and 2012 in medical journals, dealing with clinical trials on lupus
  - drugs.csv: list of drug names (3rd part of the study)
- Lupus
  - also known as systemic lupus erythematosus (SLE)
  - an autoimmune disease causing various mild to severe troubles (fever, hair loss, mouth ulceres, feeling tired, painfull and swollen joints, chest pain, swollen lymph nodes…) on tissues in many parts of the body

Text Mining - Exam subject (2)

- Variables: 7 columns in the csv file
  - Id, a number identifying the article
  - Title, the title of the article (text data, out of your study)
  - **Year**, the year of publication (categorical data used for agregation)
  - Journal, the journal in which the article has been published (categorical data)
  - First_author, the first author (categorical data)
  - **Abstract**, the abstract of the article (text data to work with)
  - Year_class, the period of publication (6 categories of year intervals)

Text Mining - Exam subject (3)

- Your work
  - Quick description of data
  - Identify more precise research topics (words which characterise the axis) and the articles which deal with these topics – using CA on non agregated data
  - Work out the evolution of research publications over years – using CA on data agregated by year
  - Work out the evolution of the drugs used over years – using CA on data agregated by year
    - Use drugs.csv to filter drug names in the vocabulary used in the abstracts