NLP
Introduction to NLP

Examples of Text Classification
Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.

- Authorship of 12 of the letters in dispute

- 1963: solved by Mosteller and Wallace using Bayesian methods
Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…

2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

Positive or negative movie review?

• unbelievably disappointing
• Full of zany characters and richly applied satire, and some great plot twists
• this is the greatest screwball comedy ever filmed
• It was pathetic. The worst part about it was the boxing scenes.
What is the subject of this article?

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...
Well-known Datasets

- 20 newsgroups
- Reuters-21578
  - Cats: grain, acquisitions, corn, crude, wheat, trade…
- WebKB
  - [http://www-2.cs.cmu.edu/~webkb/](http://www-2.cs.cmu.edu/~webkb/)
  - course, student, faculty, staff, project, dept, other
- RCV1
  - [http://www.daviddlewis.com/resources/testcollections/rcv1/](http://www.daviddlewis.com/resources/testcollections/rcv1/)
  - Larger Reuters corpus