### **Outline**

- Background
  - Wikipedia Miner
  - Wikipedia-based measures
- Exercise
  - Lexical Matching on Anchor Text

# Wikipedia Miner [Milne & Witten 2008b]

- Open source
- (Public) web service
  - Java
  - Hadoop preprocessing pipeline
- Lexical matching + machine learning
- See <a href="http://wikipedia-miner.cms.waikato.ac.nz">http://wikipedia-miner.cms.waikato.ac.nz</a>



### wikipediaminer

help demos services home

#### introduction

the demos search compare annotate

### Text to annotate

Wikipedia is a free, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its name is a portmanteau of the words wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning 'fast') and encyclopedia.

Wikipedia's 12 million articles (2.77 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in January 2001 by Jimmy Wales and Larry Sanger, it is currently the most popular general reference work on the Internet.

show options

**Annotaate** 

### Annotated text

### MediaWiki Markup | Detected Topics

[[Wikipedia]] is a free, multilingual encyclopedia project supported by the [[Nonprofit organization|non-profit]] [[Wikimedia Foundation]]. Its name is a [[portmanteau]] of the words [[wiki]] (a technology for creating collaborative websites, from the [[Hawaiian language | Hawaiian word]] wiki, meaning 'fast') and

no definition available

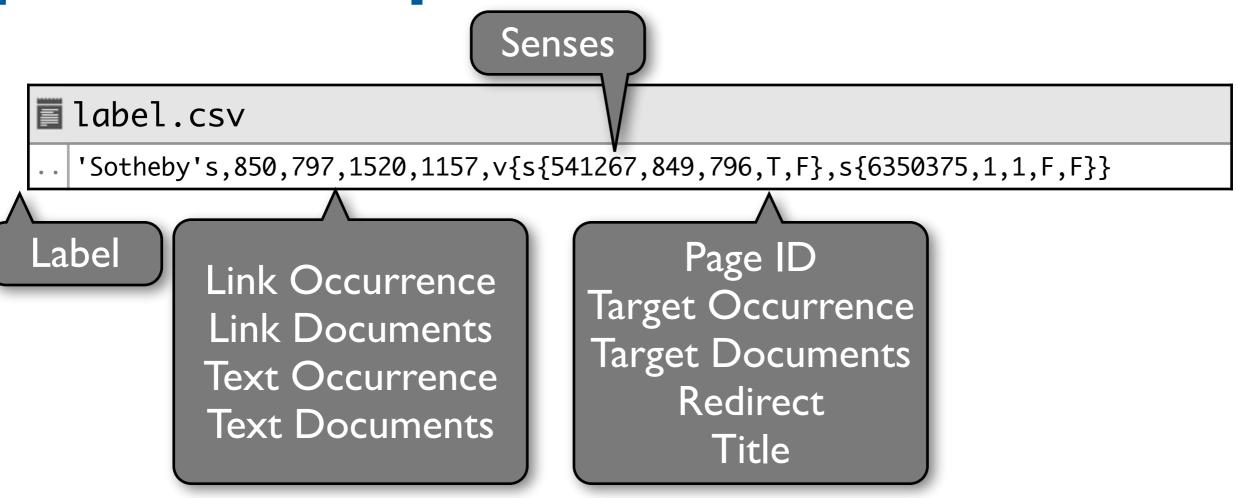
Wikipedia's 12 mill 59% probability of being a link been written collab

sh) have

and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in January 2001 by [[Jimmy Wales]] and [[Larry Sanger]], it is currently the most popular general reference work on the Internet.

## Wikipedia Miner: CSV Summary Files

[Milne & Witten 2008b]



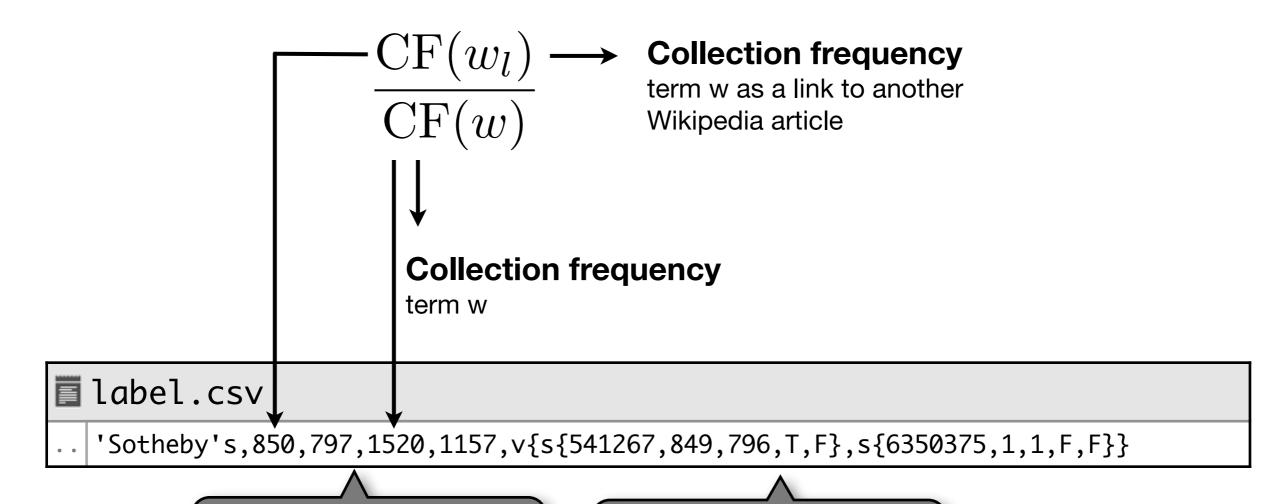
- page.csv
- 541267, 'Sotheby's, 0, 6
- .. 6350375, 'Sotheby's International Realty, 0, 6

Page ID

Title

### Wikipedia-based measures

**Keyphraseness [Mihalcea & Csomai 2007]** 

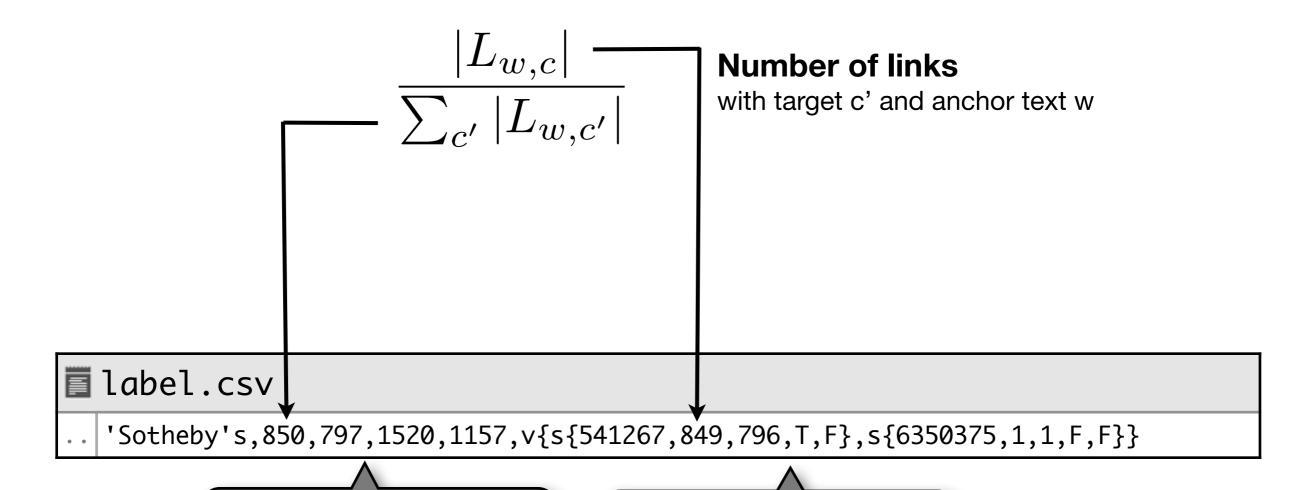


Link Occurrence
Link Documents
Text Occurrence
Text Documents

Page ID
Target Occurrence
Target Documents
Redirect
Title

### Wikipedia-based measures

**Commonness [Medelyan et al. 2008]** 



Link Occurrence
Link Documents
Text Occurrence
Text Documents

Page ID
Target Occurrence
Target Documents
Redirect
Title

### http://bit.ly/ELR-course

Codecademy

Learn

Teach

### **Entity Linking and Retrieval**

Course written by Daan Odijk

This course is created as a tutorial for Entity Linking and Retrieval. It covers web services, evaluation and fundamental methods. This Codeacademy course is part of a tutorial created for WWW2013 and SIGIR2013, see: http://bit.ly/ELR-slides.

#### 1. Entity Linking using Web Services

In this first section, we'll create a page were we can enter a text and then perform entity linking on this text. For this, we will use the DBpedia Spotlight web service.

#### 2. Evaluating Entity Linking

In this section we will evaluate an entity linking webservice using a standard dataset. We will evaluate the DBpedia Spotlight Webservice on the AQUAINT dataset.

#### 3. Entity Linking: Lexical matching of anchor text

In this exercise we will build the basic matching for lexical matching of anchor text from the ground up. This exercise is based on the University of Amsterdam Semanticizer and the University of Waikato Wikipedia Miner.

#### 4. Entity Retrieval

In this section we will evaluate Entity Retrieval using Freebase.