

David **Milne** | Ian H. **Witten**

Learning to **Link**

with

Wikipedia



The University of Waikato | **New Zealand**

Motivation

- Links between Wikipedia articles provide
 - Explanation
 - Investigation
 - Serendipity
- Can we add the same links to all documents?



David **Milne** | Ian H. **Witten**

Learning to Link

with

Wikipedia



The University of Waikato | New Zealand

Related Work

Mihalcea, R. and Csomai, A.

Wikify! linking documents to encyclopedic knowledge.

In *Proceedings of CIKM'07*, Lisbon, Portugal

INEX Link to the Wiki Track



Algorithm

Learning to Link with Wikipedia

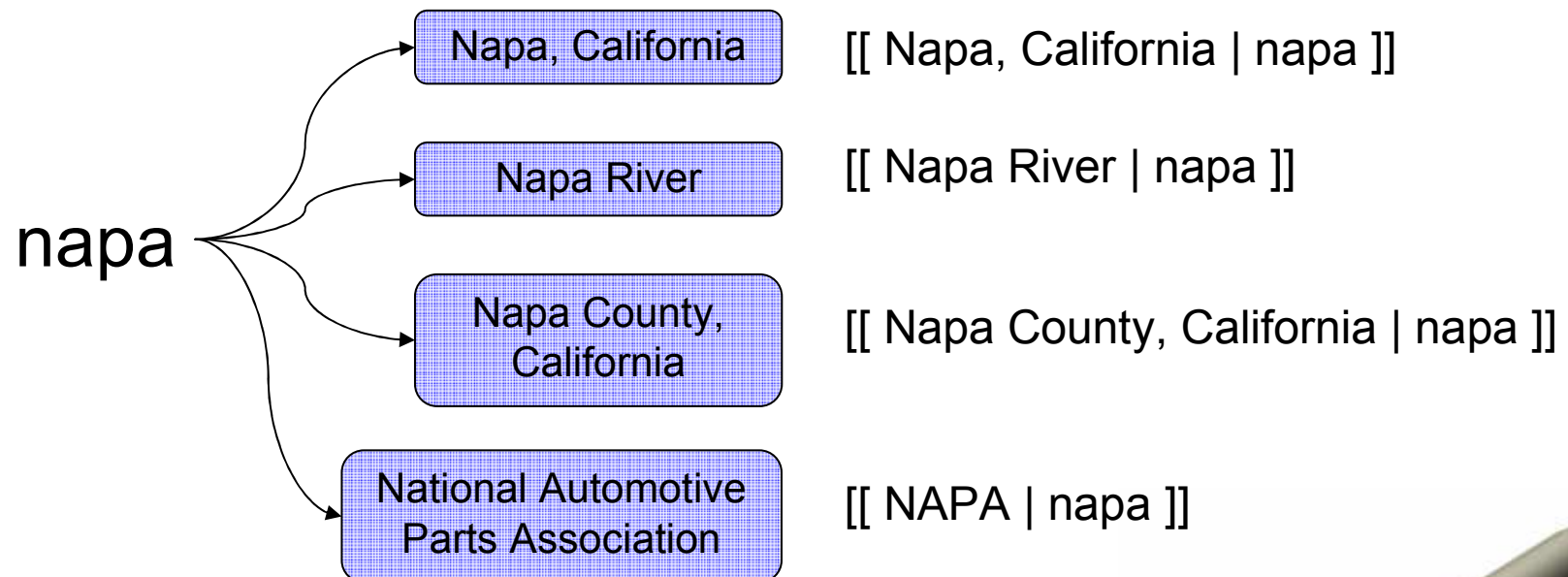
A two step process

- Link Disambiguation
- Link Selection



Algorithm | Disambiguation

For every link in Wikipedia, a human author has manually chosen the correct destination



Algorithm | Disambiguation

For every link in Wikipedia, a human author has manually chosen the correct destination

A machine-learned approach
with two main features

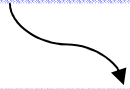
- Commonness (or prior probability)
- Relatedness to context



Algorithm | Disambiguation

Commonness

“Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”



The Global Economy	96%
Globalization	4%



Algorithm | Disambiguation

Relatedness

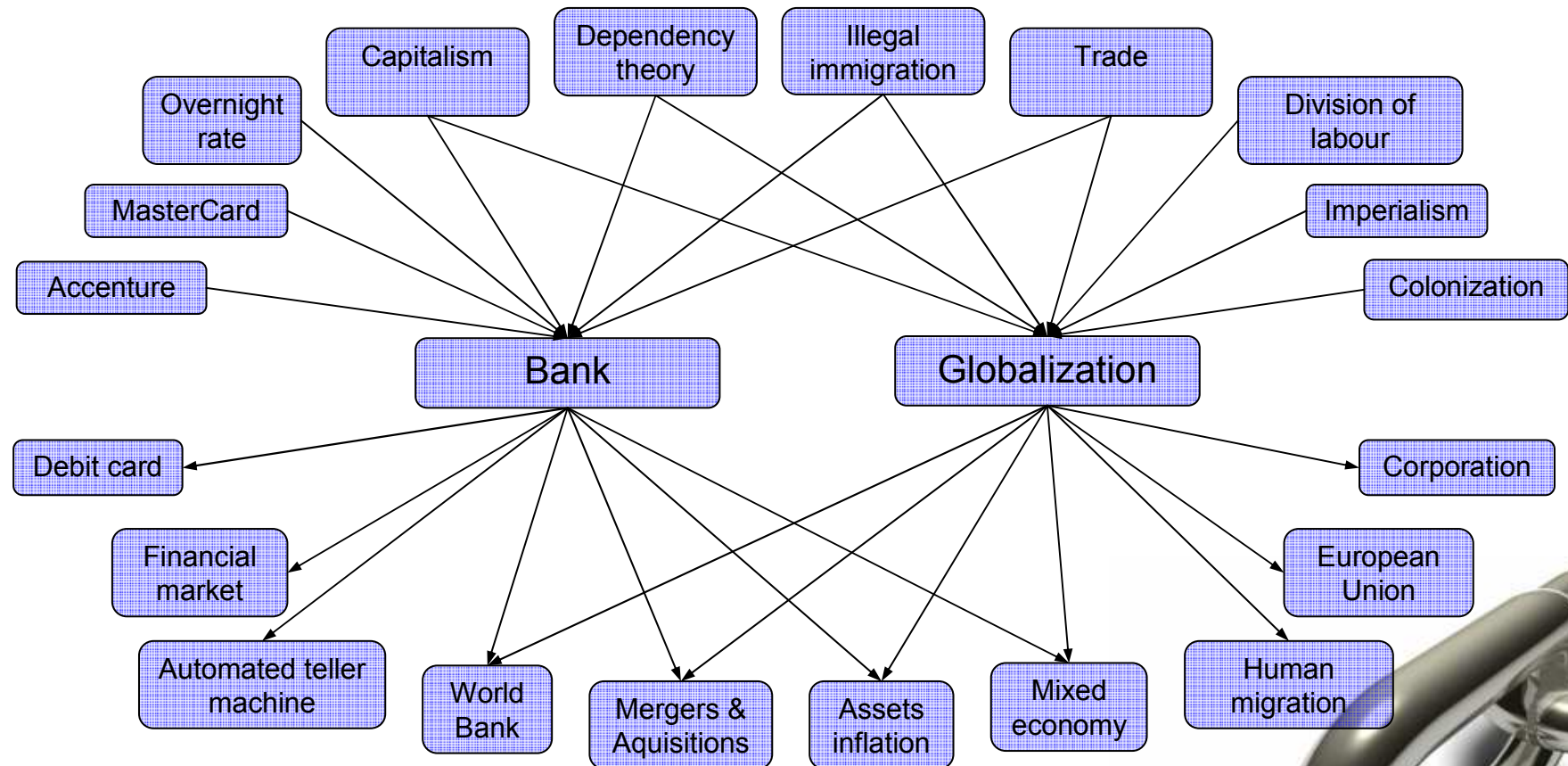
Financial institution	97.0%	0.0%
Edge of river or stream	1.8%	70.6%
An underwater hill	0.3%	2.4%
A movement in flight	0.3%	0.0%

“The story begins on the banks of the Rio Negro in the Central Amazon. A party of scientists is embarking on a voyage which they hope will provide answers to a five hundred year old mystery.”



Algorithm | Disambiguation

Relatedness



Algorithm | Disambiguation

Balancing commonness and relatedness

- Homogenous, plentiful context
 - ▲ relatedness ▼ commonness
- Ambiguous, sparse context
 - ▼ relatedness ▲ commonness
- Third feature: quality of context



Evaluation | Disambiguation

Wikipedia provides ground truth as well as training data

- trained on 500 articles
- developed and tweaked on 100 articles
- tested on 100 articles

recall **96%** precision **98%**



Algorithm | Link Selection

Napa, California

From Wikipedia, the free encyclopedia

Coordinates:  38.304764, -122.298885

"Napa" redirects here. For other uses, see [Napa \(disambiguation\)](#).

Napa is the [county seat](#) of [Napa County, California](#). It is the principal city of the Napa county [Metropolitan Statistical Area](#), which encompasses Napa county. As of the 2000 census, the city had a total population of 72,585. The area was settled in the 1830s. It was incorporated as a city in 1872.

On December 31, 2005, the [Napa River](#) overflowed and flooded the entire downtown area and thousands of acres all over Napa County. More than 4,000 residents were evacuated and 1,000 homes were flooded or destroyed.^[1] The flood of 12/31/05 was the 23rd most serious flood on the Napa River on record since 1865.

Napa, California



Location in [Napa County](#) and the state of [California](#)

Coordinates:  38°18'17"N 122°17'56"W

Algorithm | Link Selection

Every Wikipedia article is an example of how to cross-reference a document with Wikipedia.

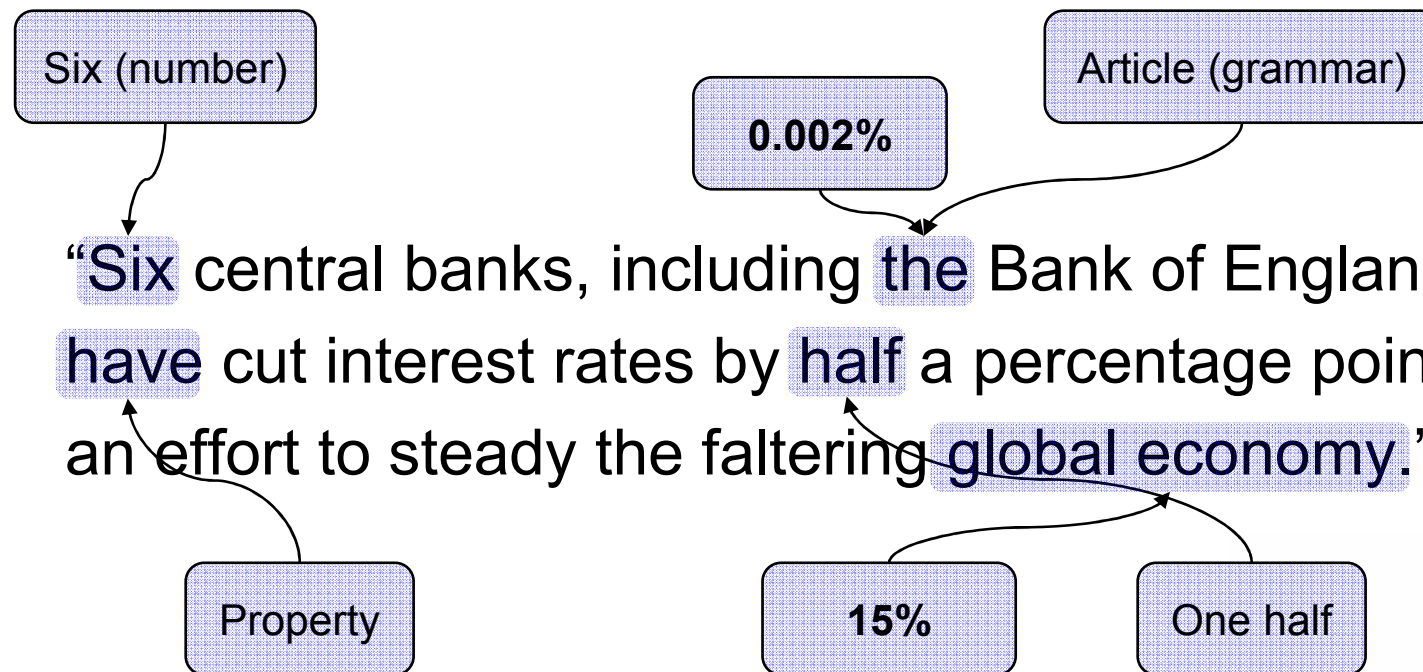
A machine-learned approach

- Detect and disambiguate every term or phrase that might be linked.
- Use features of concepts and where they are found to learn what to link.



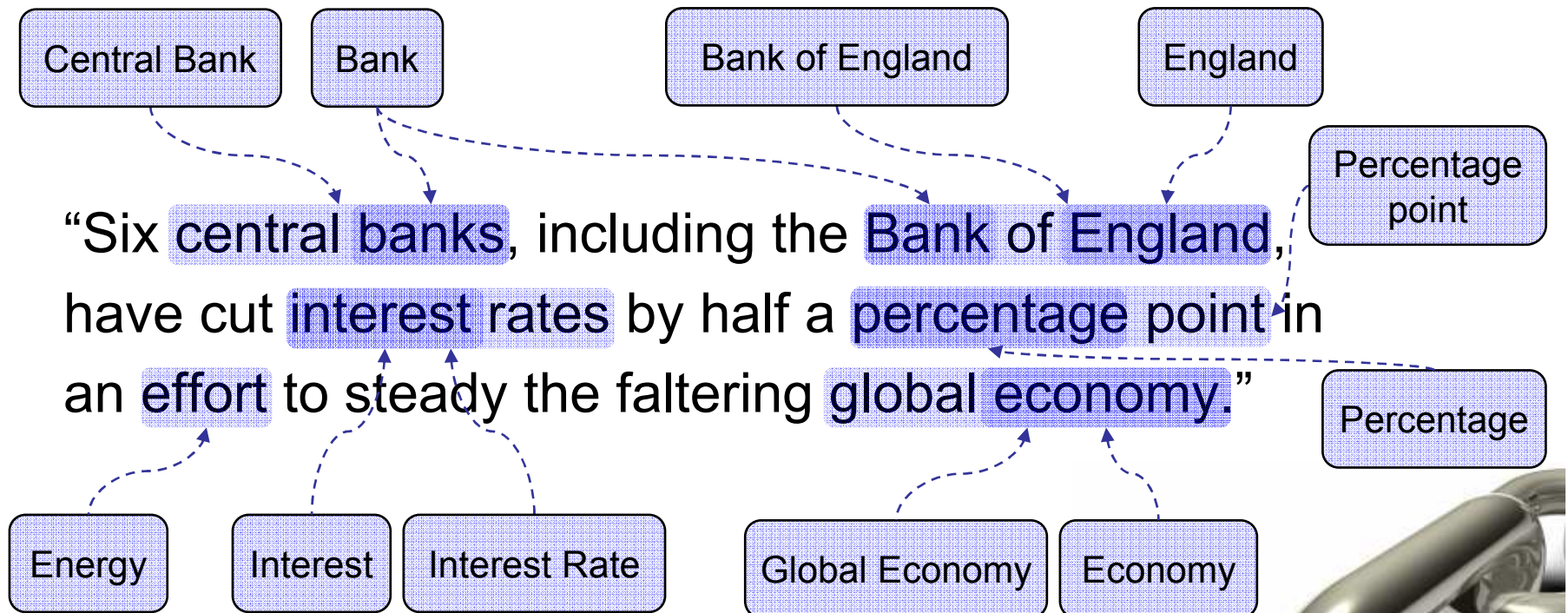
Algorithm | Link Selection

Wikipedia's links provide a huge vocabulary of which terms correspond to concepts



Algorithm | Link Selection

Wikipedia's links provide a huge vocabulary of which terms correspond to concepts



Algorithm | Link Selection

Features

- Link Probability
- Relatedness
- Disambiguation Confidence
- Generality
- Location and Spread



Evaluation | Link Selection

On 100 randomly selected Wikipedia articles

recall **74%** precision **74%**

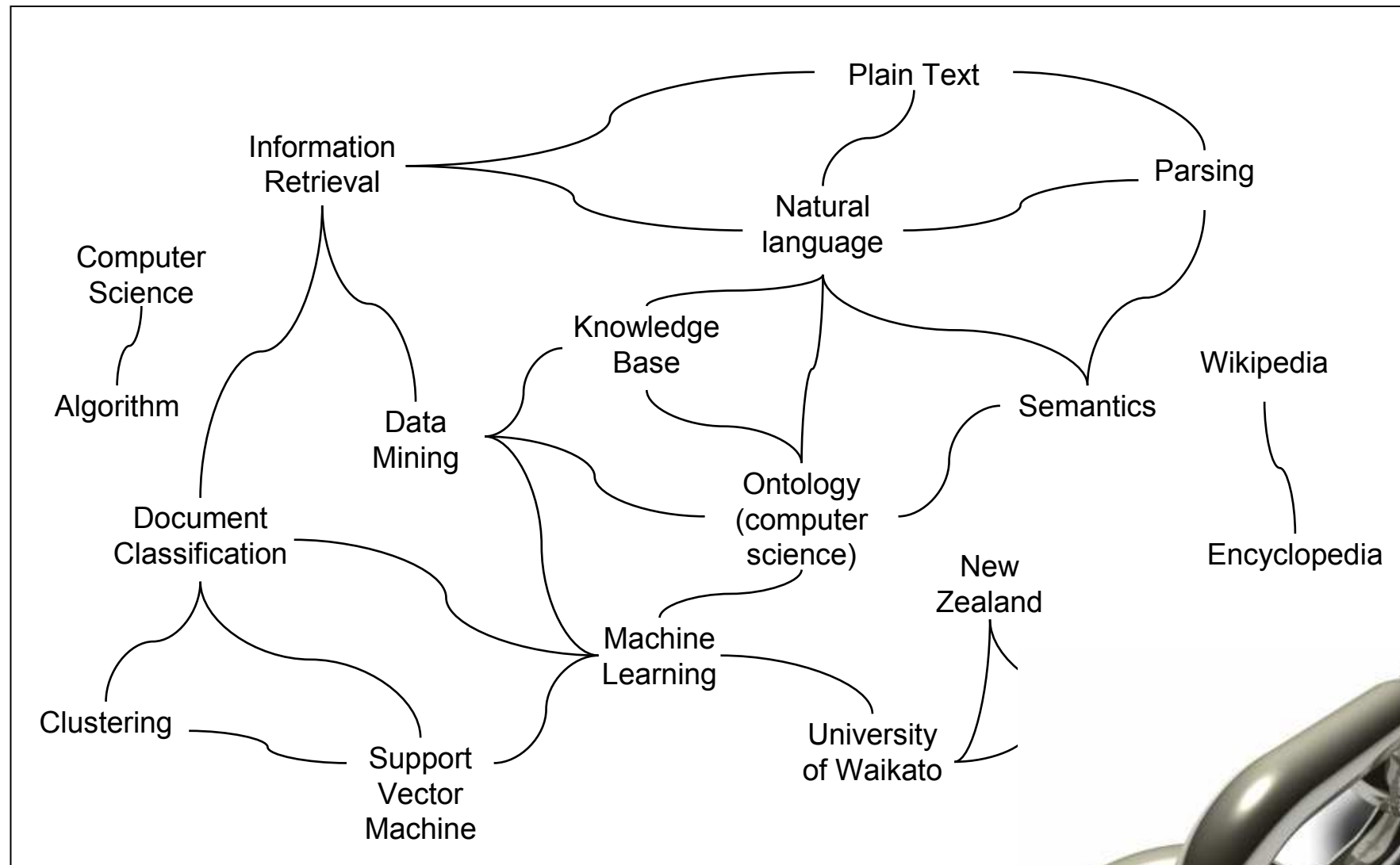
On 50 news documents, with human judgments

recall **73%** precision **76%**

50% improvement on previous work



Implications | and applications



Implications | and applications

We can...

...add explanatory links to any document

- Augment news stories, blogs, educational materials
- Assist creation of new Wikipedia articles

...improve how documents are represented

- Information Retrieval
- Topic Indexing (Olena Medelyan)
- Document Clustering (Anna Huang)
- Multi-document Summarization (Vivi Nastase)



Thanks! | Any Questions?

www.nzdl.org/wikification

