

# Word Hyphenation Correction and Query Expansion using Wikipedia for Book Retrieval

Romain Deveaud, Florian Boudin, Eric SanJuan and Patrice Bellot  
LIA - University of Avignon

INEX'10 Book Track, December 13-15

# Introduction

- Digitizing books with Optical Character Recognition (OCR)
  - Gutenberg
  - Google Books

# Introduction

- Digitizing books with Optical Character Recognition (OCR)
  - Gutenberg
  - Google Books
- Hyphenated words

Too often people are prepared to accept that alcoholic **bever-**  
**ages** are thoroughly bad: a portal to doom, debauchery, and  
damnation...

} indexed as **bever-** and **ages**

# Introduction

- Digitizing books with Optical Character Recognition (OCR)
  - Gutenberg
  - Google Books

- Hyphenated words

Too often people are prepared to accept that alcoholic **bever-**  
**ages** are thoroughly bad: a portal to doom, debauchery, and  
damnation...

} indexed as **bever-** and **ages**

- Query Expansion

# Introduction

- Digitizing books with Optical Character Recognition (OCR)
  - Gutenberg
  - Google Books

- Hyphenated words

Too often people are prepared to accept that alcoholic **bever-**  
**ages** are thoroughly bad: a portal to doom, debauchery, and  
damnation...

} indexed as **bever-** and **ages**

- Query Expansion
  - Pseudo-relevance feedback
    - Wikipedia for selecting **expansion words**
    - One **Wikipedia** page for **each query** (Koolen *et al.*, WSDM'09)

# Introduction

- Digitizing books with Optical Character Recognition (OCR)
  - Gutenberg
  - Google Books

- Hyphenated words

Too often people are prepared to accept that alcoholic **bever-**  
**ages** are thoroughly bad: a portal to doom, debauchery, and  
damnation...

} indexed as **bever-** and **ages**

- Query Expansion
  - Pseudo-relevance feedback
    - Wikipedia for selecting **expansion words**
    - One **Wikipedia** page for **each query** (Koolen *et al.*, WSDM'09)
  - MRF model (Metzler and Croft, SIGIR'05)

# Word Hyphenation Correction

- For each couple of lines (L1, L2) of each book

L1: Too often people are prepared to accept that alcoholic w1[bever]-  
L2: w2[ages] are thoroughly bad: a portal to doom, debauchery... } concat(w1,w2) : beverages

# Word Hyphenation Correction

- For each **couple of lines** (**L1**, **L2**) of each book

L1: Too often people are prepared to accept that alcoholic **w1[bever]-**  
L2: **w2[ages]** are thoroughly bad: a portal to doom, debauchery... } concat(**w1,w2**) : **beverages**

- English Gigaword lexicon (118,221 unique words)



# Word Hyphenation Correction

- For each **couple of lines** (**L1**, **L2**) of each book

L1: Too often people are prepared to accept that alcoholic **w1[bever]-**  
L2: **w2[ages]** are thoroughly bad: a portal to doom, debauchery... } concat(**w1,w2**) : **beverages**

- English Gigaword lexicon (118,221 unique words)
- Indexing and retrieval tasks performed using **Indri**<sup>1</sup>

# Word Hyphenation Correction

- For each **couple of lines** (**L1**, **L2**) of each book

L1: Too often people are prepared to accept that alcoholic **w1[bever]-**  
L2: **w2[ages]** are thoroughly bad: a portal to doom, debauchery... } concat(**w1,w2**) : **beverages**

- English Gigaword lexicon (118,221 unique words)
- Indexing and retrieval tasks performed using **Indri**<sup>1</sup>
  - Porter stemmer

# Word Hyphenation Correction

- For each **couple of lines** (**L1**, **L2**) of each book

L1: Too often people are prepared to accept that alcoholic **w1[bever]-**  
L2: **w2[ages]** are thoroughly bad: a portal to doom, debauchery... } concat(w1,w2) : **beverages**

- English Gigaword lexicon (118,221 unique words)
- Indexing and retrieval tasks performed using **Indri**<sup>1</sup>
  - Porter stemmer
  - Language modeling (LM), Dirichlet smoothing

# Word Hyphenation Correction

- For each **couple of lines** (**L1**, **L2**) of each book

L1: Too often people are prepared to accept that alcoholic **w1[bever]-**  
L2: **w2[ages]** are thoroughly bad: a portal to doom, debauchery... } concat(**w1,w2**) : **beverages**

- English Gigaword lexicon (118,221 unique words)
- Indexing and retrieval tasks performed using **Indri**<sup>1</sup>
  - Porter stemmer
  - Language modeling (LM), Dirichlet smoothing
  - 100 books retrieved for each query

# Word Hyphenation Correction

Table 1. Book retrieval results on both initial and corrected INEX 2009 Book Track corpus in terms of Mean Average Precision (MAP) and precision at 10 (P@10).

Model	Uncorrected data		Corrected data	
	MAP	P@10	MAP	P@10
LM, $\mu = 2500$	0.302	0.486	0.304	0.507
LM, $\mu = 1000$	0.299	0.493	0.302	0.507
LM, $\mu = 0$	0.244	0.443	0.243	0.450

# Baselines

- Two **simple** models

# Baselines

- Two **simple** models
- Bag of words
  - **Language modeling** approach
  - Dirichlet smoothing,  $\mu = 2000$

# Baselines

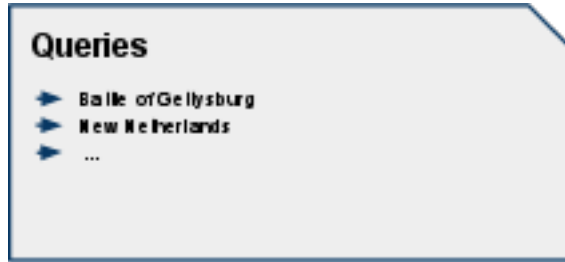
- Two **simple** models
- Bag of words
  - **Language modeling** approach
  - Dirichlet smoothing,  $\mu = 2000$
- `baseline_1` : content of `<query>` fields



# Baselines

- Two **simple** models
- Bag of words
  - **Language modeling** approach
  - Dirichlet smoothing,  $\mu = 2000$
- `baseline_1` : content of `<query>` fields
- `baseline_2` : content of `<fact>` fields

# Query Expansion using Wikipedia



**Wikipedia**

# Query Expansion using Wikipedia

**Queries**

- ▶ Battle of Gettysburg
- ▶ New Netherlands
- ▶ ...

[http://en.wikipedia.org/wiki/Battle\\_of\\_Gettysburg](http://en.wikipedia.org/wiki/Battle_of_Gettysburg)

## Battle of Gettysburg

The Battle of Gettysburg (locally /'ɡɛtɪsbɜːrɡ/ (listen), with an ss sound), fought July 1–3, 1863, in and around the town of Gettysburg, Pennsylvania, was the battle with the largest number of casualties in the American Civil War [6] and is often described as the war's turning point.[7] Union Maj. Gen. George Gordon Meade's Army of the Potomac defeated attacks by Confederate Gen. Robert E. Lee's Army of Northern Virginia, ending Lee's invasion of the North.

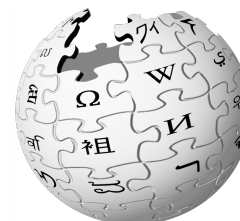
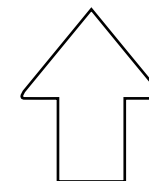
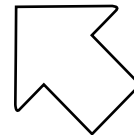
---

[http://en.wikipedia.org/wiki/New\\_Netherlands](http://en.wikipedia.org/wiki/New_Netherlands)

## New Netherland

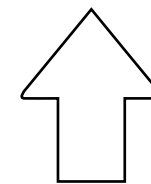
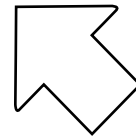
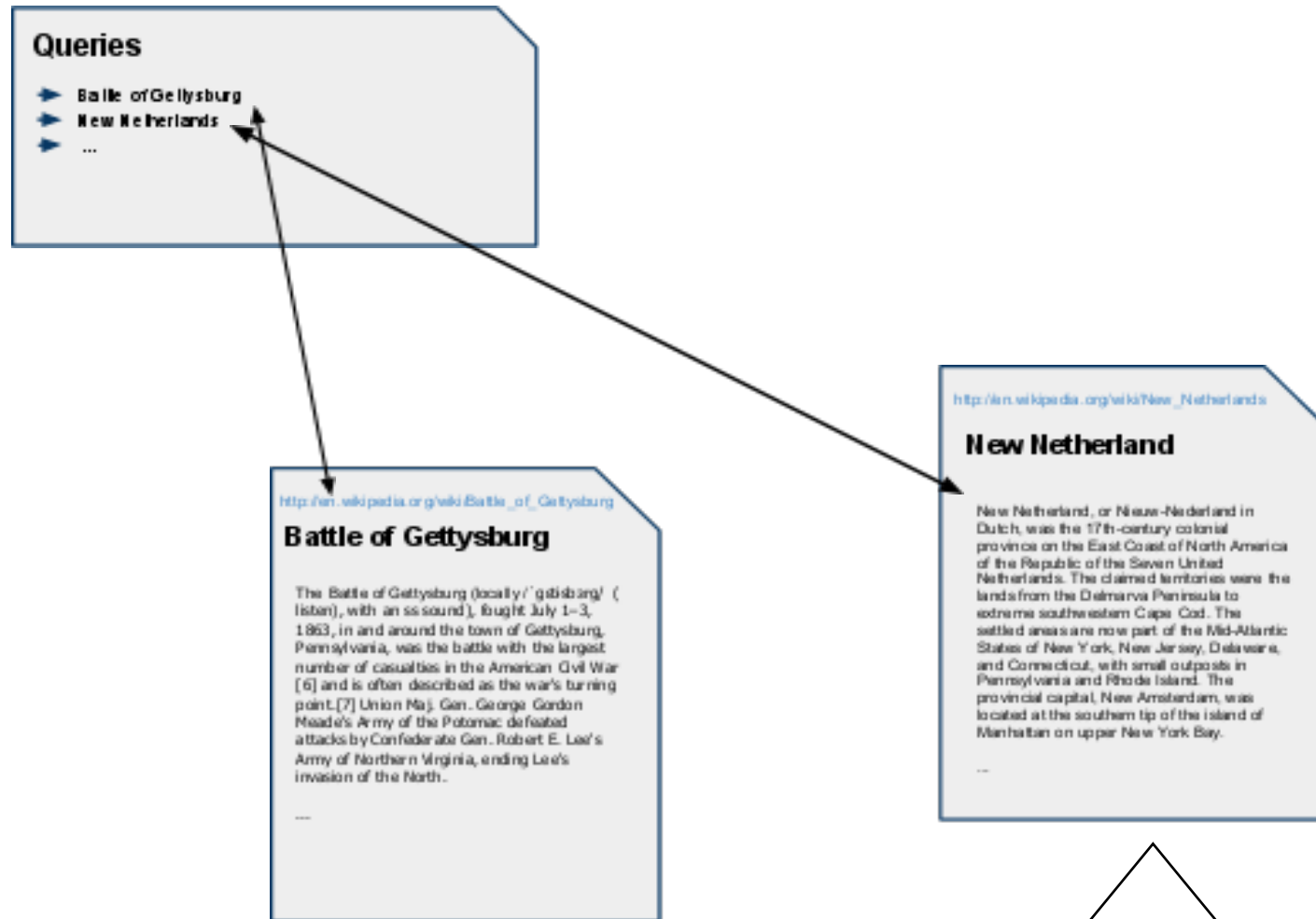
New Netherland, or *Nieuw-Nederland* in Dutch, was the 17th-century colonial province on the East Coast of North America of the Republic of the Seven United Netherlands. The claimed territories were the lands from the Delaware Peninsula to extreme southwestern Cape Cod. The settled areas are now part of the Mid-Atlantic States of New York, New Jersey, Delaware, and Connecticut, with small outposts in Pennsylvania and Rhode Island. The provincial capital, New Amsterdam, was located at the southern tip of the island of Manhattan on upper New York Bay.

---



**Wikipedia**

# Query Expansion using Wikipedia



**Wikipedia**

# Query Expansion using Wikipedia

Koolen *et al.* (WSDM'09):

battle of gettysburg

# Query Expansion using Wikipedia

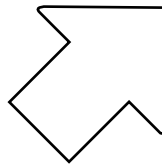
Koolen *et al.* (WSDM'09):

battle of gettysburg  
 $t_1 \dots t_N$

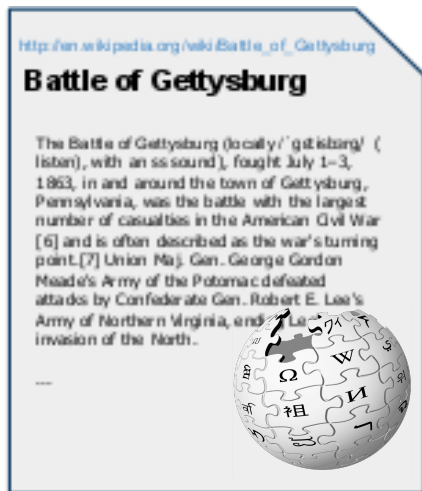
# Query Expansion using Wikipedia

Koolen *et al.* (WSDM'09):

battle of gettysburg  
 $t_1 \dots t_N$   
N top-ranked words



tf.idf



# Query Expansion using Wikipedia

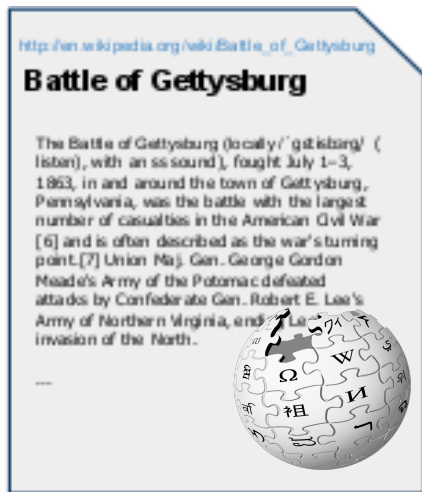
Koolen *et al.* (WSDM'09):

battle of gettysburg

$t_1 \dots t_N$

**N top-ranked words**

**tf** idf





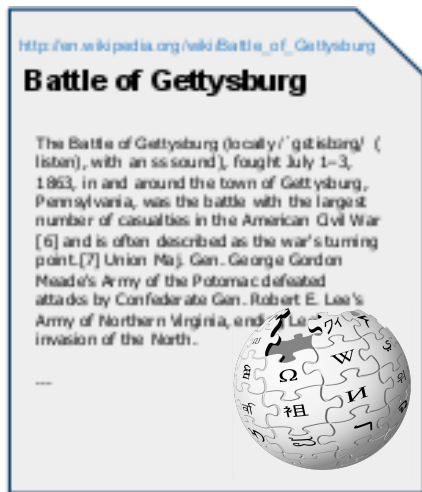
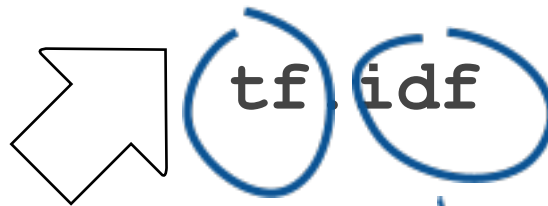
# Query Expansion using Wikipedia

Koolen *et al.* (WSDM'09):

battle of gettysburg

$t_1 \dots t_N$

**N top-ranked words**



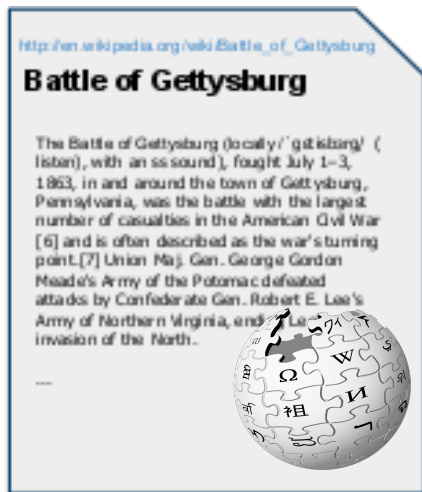
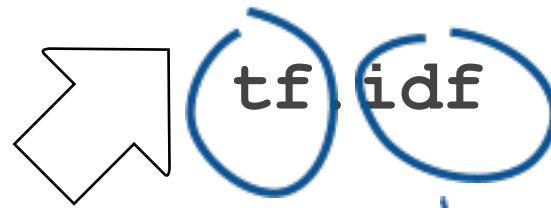
INEX Book Track corpus

# Query Expansion using Wikipedia

Koolen *et al.* (WSDM'09):

```
#weight ( N #combine ( battle of gettysburg )  
1 #weight ( t1 ... tN ) )
```

N top-ranked words



INEX Book Track corpus

# Query Expansion using Wikipedia


```
#weight ( N #combine ( battle of gettysburg )  
1 #weight ( t1 ... tN ) )
```

N top-ranked words

tf, idf

[http://en.wikipedia.org/wiki/Battle\\_of\\_Gettysburg](http://en.wikipedia.org/wiki/Battle_of_Gettysburg)  
**Battle of Gettysburg**

The Battle of Gettysburg (localy /ˈɡɛtɪzburɪ/ (listen), with an ss sound), fought July 1–3, 1863, in and around the town of Gettysburg, Pennsylvania, was the battle with the largest number of casualties in the American Civil War [6] and is often described as the war's turning point.[7] Union Maj. Gen. George Gordon Meade's Army of the Potomac defeated attacks by Confederate Gen. Robert E. Lee's Army of Northern Virginia, ending Lee's invasion of the North.



entropy

$$E(t_i) = -tf_{t_i} \times \log_2(tf_{t_i})$$



INEX Book Track corpus

# Query Expansion using Wikipedia

score normalization

#weight ( N #combine ( battle of gettysburg )  
 1 #weight (  $w_1 t_1 \dots w_N t_N$  ) )

N top-ranked words


tf, idf

[http://en.wikipedia.org/wiki/Battle\\_of\\_Gettysburg](http://en.wikipedia.org/wiki/Battle_of_Gettysburg)

**Battle of Gettysburg**

The Battle of Gettysburg (localy /'gɛtɪzɪŋ/ (listen), with an ss sound), fought July 1–3, 1863, in and around the town of Gettysburg, Pennsylvania, was the battle with the largest number of casualties in the American Civil War [6] and is often described as the war's turning point.[7] Union Maj. Gen. George Gordon Meade's Army of the Potomac defeated attacks by Confederate Gen. Robert E. Lee's Army of Northern Virginia, ending Lee's invasion of the North.

...



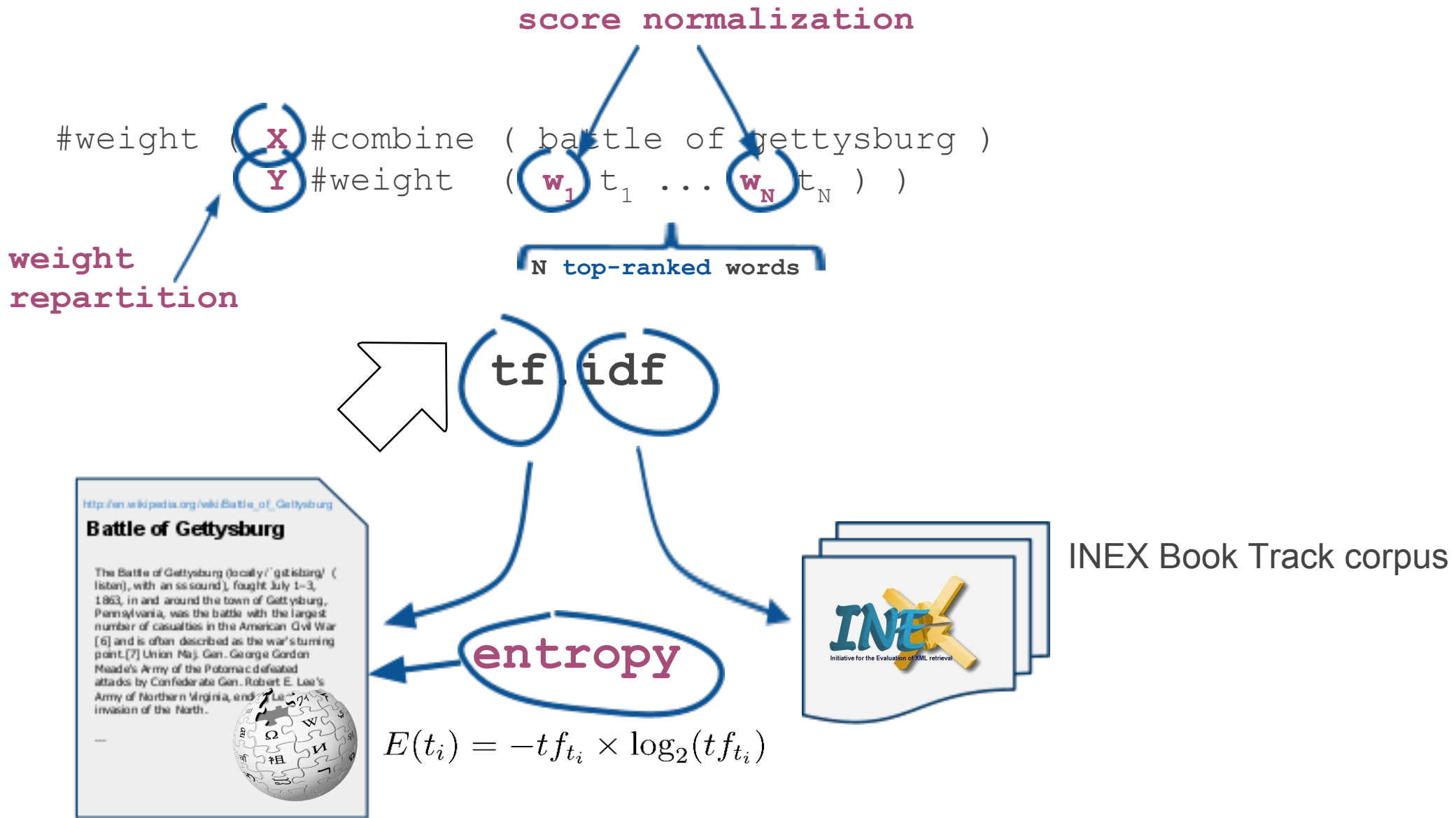
entropy

$$E(t_i) = -tf_{t_i} \times \log_2(tf_{t_i})$$

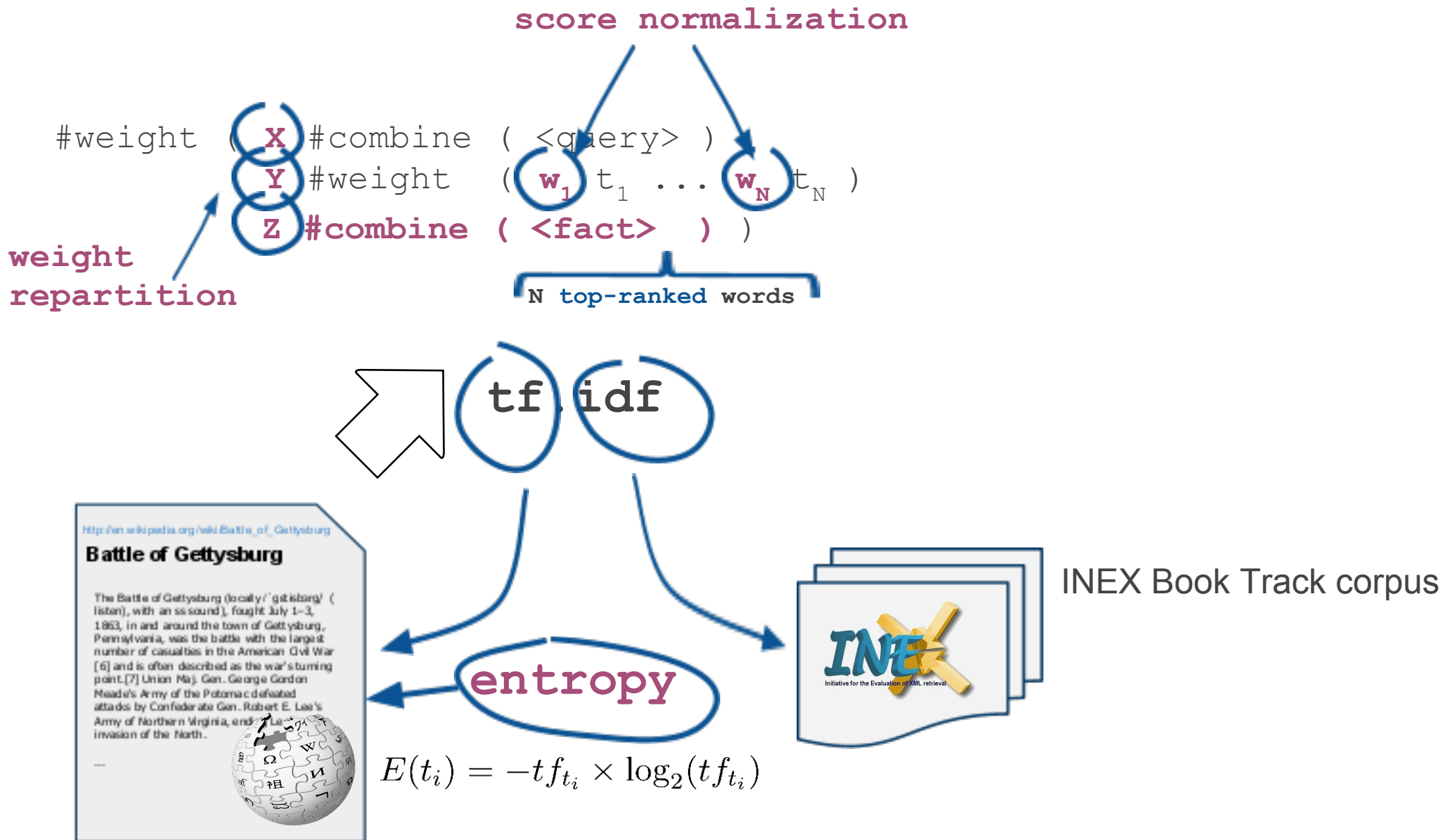


INEX Book Track corpus

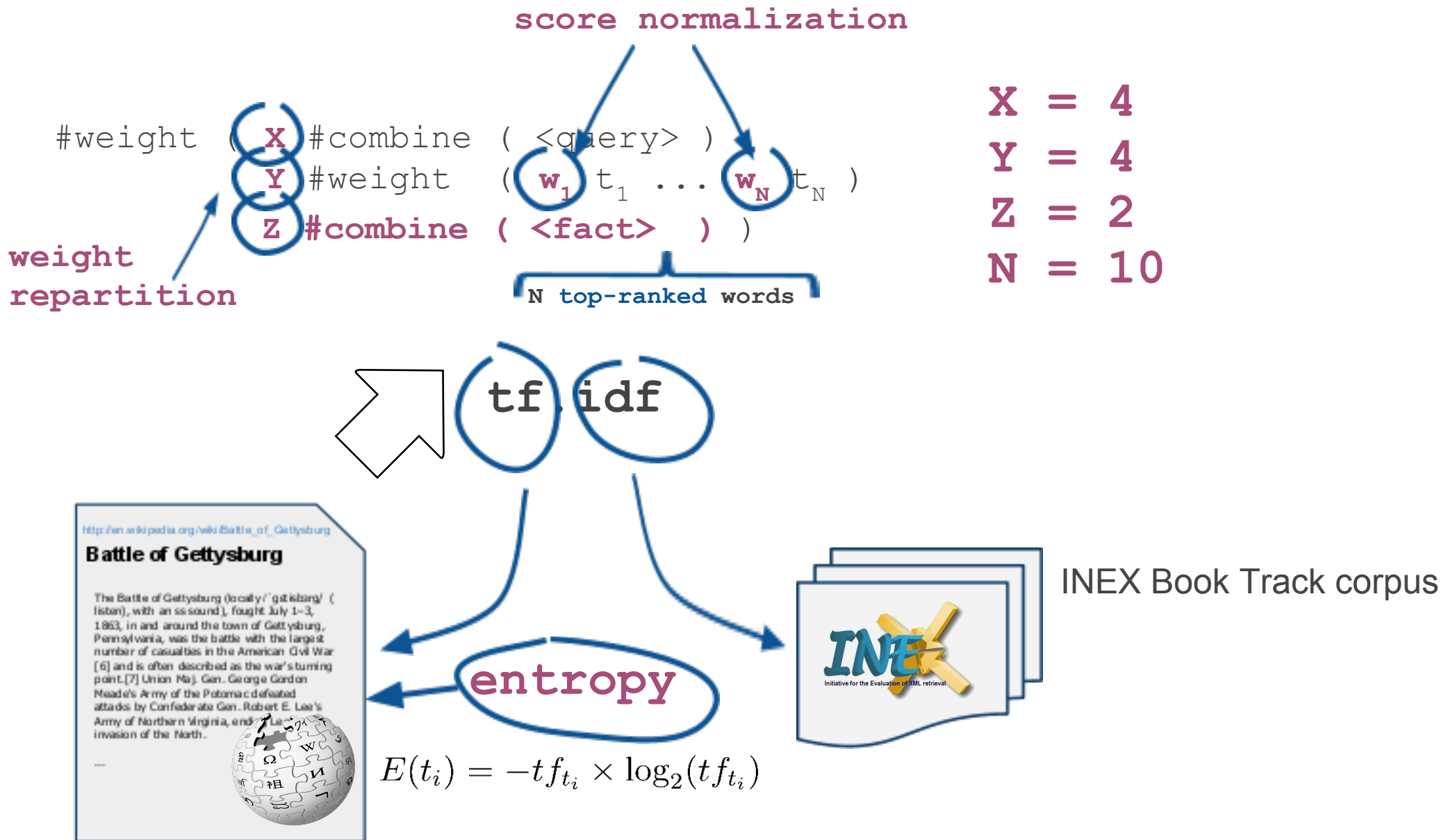
# Query Expansion using Wikipedia



# Query Expansion using Wikipedia



# Query Expansion using Wikipedia



# Query Expansion using Stanford Parser

- Multiword phrases are detected using the [Stanford parser](http://nlp.stanford.edu/software/lex-parser.shtml)<sup>1</sup>

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>



# Query Expansion using Stanford Parser

- Multiword phrases are detected using the [Stanford parser](#)<sup>1</sup>
  - "New York" instead of "New" and "York"

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

# Query Expansion using Stanford Parser

- Multiword phrases are detected using the [Stanford parser](#)<sup>1</sup>
  - "New York" instead of "New" and "York"
- Combination of several features
  - single terms ([unigrams](#))
  - exact phrases (words appearing in [sequence](#))
  - [unordered](#) windows ([no exact sequence order](#))

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

# Query Expansion using Stanford Parser

- Multiword phrases are detected using the [Stanford parser](#)<sup>1</sup>
  - "New York" instead of "New" and "York"
- Combination of several [features](#)
  - single terms ([unigrams](#))
  - exact phrases (words appearing in [sequence](#))
  - [unordered](#) windows ([no exact sequence order](#))
- Feature weights follows the author's<sup>2</sup> recommendation

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

<sup>2</sup>D. Metzler and W. B. Croft (SIGIR'05)

# Query Expansion using Stanford Parser

- Multiword phrases are detected using the [Stanford parser](#)<sup>1</sup>
  - "New York" instead of "New" and "York"
- Combination of several [features](#)
  - single terms ([unigrams](#))
  - exact phrases (words appearing in [sequence](#))
  - [unordered](#) windows ([no exact sequence order](#))
- Feature weights follows the author's<sup>2</sup> recommendation
- We only use the `<fact>` fields

<sup>1</sup><http://nlp.stanford.edu/software/lex-parser.shtml>

<sup>2</sup>D. Metzler and W. B. Croft (SIGIR'05)

# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval

# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval
- Two book retrieval **baselines**

# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval
- Two book retrieval **baselines**
- Two **Query Expansion** approaches

# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval
- Two book retrieval **baselines**
- Two **Query Expansion** approaches
  - using **Wikipedia** for term extraction



# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval
- Two book retrieval **baselines**
- Two **Query Expansion** approaches
  - using **Wikipedia** for term extraction
    - entropy
    - tf.idf

# Conclusion and Future Work

- Evaluation of hyphenation correction impact on **focused search** tasks
  - page retrieval
  - extent retrieval
- Two book retrieval **baselines**
- Two **Query Expansion** approaches
  - using **Wikipedia** for term extraction
    - entropy
    - tf.idf
  - using the **Stanford parser** with MRFs

Thank you for your attention

# Appendix

**Table 2.** Query expansion performance at the top N words ranked by *tf.idf* or *entropy*, on the corrected and the uncorrected corpus, with (X:Y) weights (<sup>†</sup> : t.test < 0.05; <sup>‡</sup>: t.test < 0.01).

Method	corpus	N = 5		N = 10		N = 20		N = 50	
		MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
<i>entropy</i> (1:3)	Cor.	0.301	0.489	0.346	0.564	0.330	0.529	0.353	0.564
<i>entropy</i> (2:2)	Cor.	0.327	0.557	0.348	0.564	0.361	0.592 <sup>†</sup>	<b>0.363</b>	<b>0.593<sup>‡</sup></b>
<i>entropy</i> (2:2)	UnC.	0.323	0.542	0.346	0.579	0.358	0.586 <sup>†</sup>	0.361	<b>0.593<sup>‡</sup></b>
<i>entropy</i> (3:1)	Cor.	0.330	0.564	0.342 <sup>†</sup>	0.564	0.349	0.564	0.347	0.557
<i>tf.idf</i> (1:3)	Cor.	0.245	0.479	0.249	0.450	0.257	0.464	0.246	0.486
<i>tf.idf</i> (2:2)	Cor.	0.277	0.486	0.290	0.521	0.289	0.140	0.295	0.514
<i>tf.idf</i> (3:1)	Cor.	0.310	0.536	0.311	0.543	<b>0.317</b>	<b>0.557</b>	0.314	0.536
<i>tf.idf</i> (3:1)	UnC.	0.307	0.529	0.311	0.543	0.314	<b>0.557</b>	0.313	0.529
Koolen <i>et al.</i>	Cor.	0.308	<b>0.550</b>	<b>0.321</b>	0.536	0.301	0.521	0.306	0.507
Koolen <i>et al.</i>	UnC.	0.308	<b>0.550</b>	0.315	0.521	0.300	0.514	0.304	0.500