Correction de césures et enrichissement de requêtes pour la recherche de livres

Romain Deveaud, Florian Boudin, Eric SanJuan, Patrice Bellot

LIA – Université d'Avignon

CORIA 2011, 16-18 Mars

Introduction

Numérisation des livres

- Reconnaissance Optique des Caractères (ROC)
- Gutenberg
- Google Books

Collection de livres INEX Book Track

```
<region>
<section id="0" key="1332" label="SEC_BODY">
  <line>primitive tradition which is recorded in our Sacred</line>
  <line>Books ? When God created the world He simply</line>
```

Introduction

Césures

```
Le pastis 51, comme d'autres bois-
sons anisées, est appelé dans le Sud de
la France : un jaune, un flan ou un flaï.
```

Indexé comme **bois-** et **sons**

Enrichissement de requêtes

Wikipedia comme base externe de connaissances

Une page Wikipedia pour chaque requête (Koolen *et al.*, WSDM'09)

Correction de césures

Pour chaque couple de lignes (`L1,L2`) de chaque livre

L1: Le pastis 51, comme d'autres **w1[**bois**]-**
L2: w2[sons**]** anisées, est appelé dans le Sud

concat(**w1**,**w2**) :
boissons

Lexique issu du corpus English Gigaword

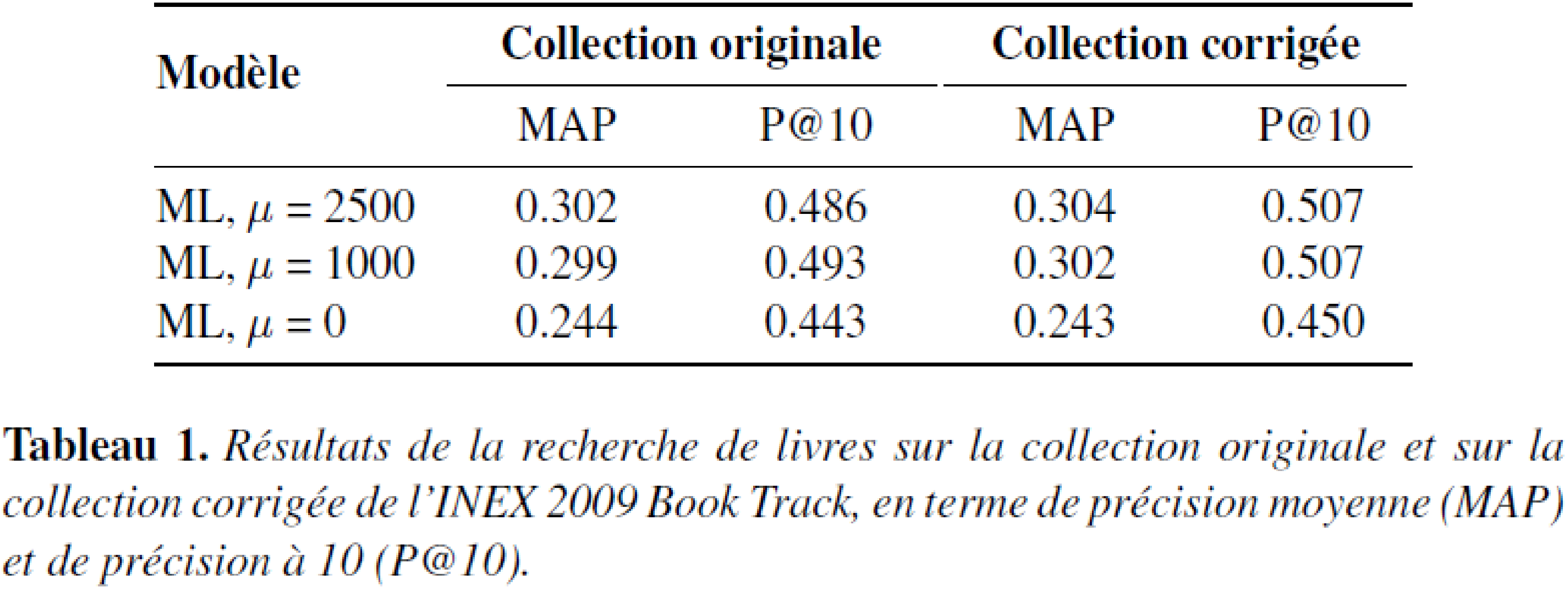
613 107 923 lignes contiennent des césures

- Et 37 551 834 sont corrigées (6,125%)

Comment évaluer?

- Indexation et extraction utilisant Indri
- Collection et requêtes (16) provenant de la Book Track INEX 2009

Correction de césures



Modèle	Collection originale		Collection corrigée	
	MAP	P@10	MAP	P@10
ML, $\mu = 2500$	0.302	0.486	0.304	0.507
ML, $\mu = 1000$	0.299	0.493	0.302	0.507
ML, $\mu = 0$	0.244	0.443	0.243	0.450

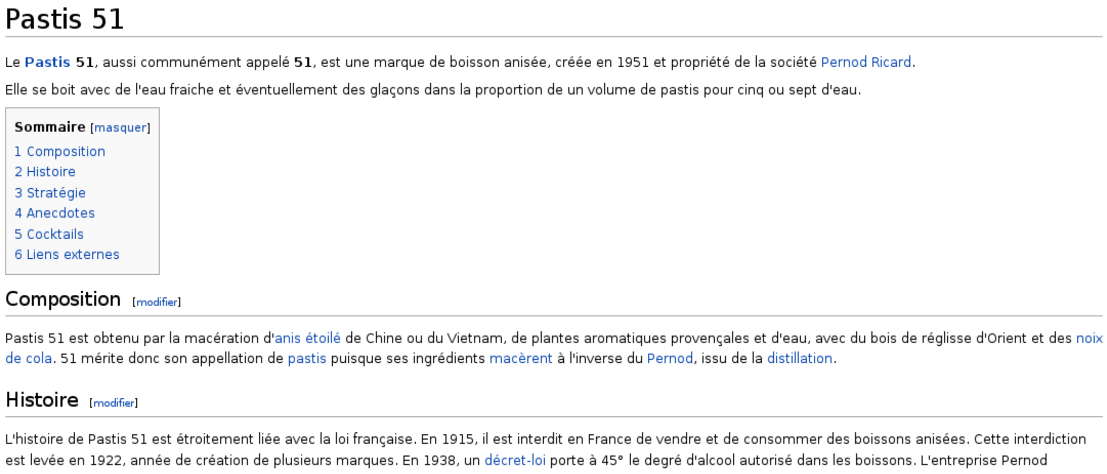
Tableau 1. *Résultats de la recherche de livres sur la collection originale et sur la collection corrigée de l'INEX 2009 Book Track, en terme de précision moyenne (MAP) et de précision à 10 (P@10).*

Enrichissement de requêtes avec Wikipédia

Koolen *et al.* (WSDM'09)

pastis 51

⟺



Pastis 51

Le **Pastis 51**, aussi communément appelé **51**, est une marque de boisson anisée, créée en 1951 et propriété de la société Pernod Ricard.

Elle se boit avec de l'eau fraiche et éventuellement des glaçons dans la proportion de un volume de pastis pour cinq ou sept d'eau.

Sommaire [masquer]

Composition [modifier]

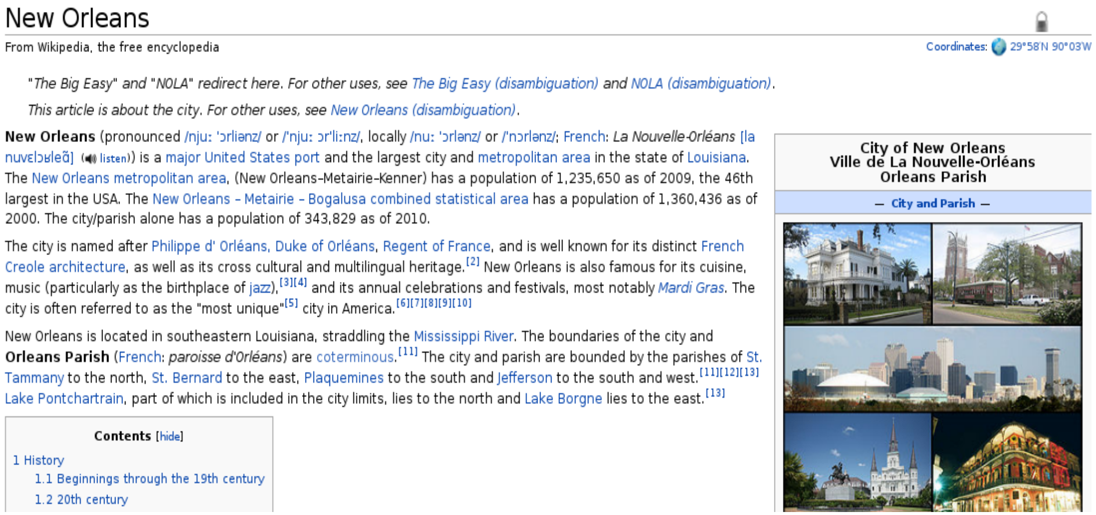
Pastis 51 est obtenu par la macération d'anis étoilé de Chine ou du Vietnam, de plantes aromatiques provençales et d'eau, avec du bois de réglisse d'Orient et des noix de cola. 51 mérite donc son appellation de pastis puisque ses ingrédients macèrent à l'inverse du Pernod, issu de la distillation.

Histoire [modifier]

L'histoire de Pastis 51 est étroitement liée avec la loi française. En 1915, il est interdit en France de vendre et de consommer des boissons anisées. Cette interdiction est levée en 1922, année de création de plusieurs marques. En 1938, un décret-loi porte à 45° le degré d'alcool autorisé dans les boissons. L'entreprise Pernod

new orleans

⟺



New Orleans

From Wikipedia, the free encyclopedia

Coordinates: 29°58′N 90°03′W

"The Big Easy" and "NOLA" redirect here. For other uses, see The Big Easy (disambiguation) and NOLA (disambiguation).

This article is about the city. For other uses, see New Orleans (disambiguation).

New Orleans (pronounced /njuː ˈɔrliənz/ or /ˈnjuː ɔrˈliːnz/, locally /nuː ˈɔrlənz/ or /ˈnɔrlənz/; French: *La Nouvelle-Orléans* [la nuvɛlɔʁleɑ̃] (listen)) is a major United States port and the largest city and metropolitan area in the state of Louisiana. The New Orleans metropolitan area, (New Orleans–Metairie–Kenner) has a population of 1,235,650 as of 2009, the 46th largest in the USA. The New Orleans – Metairie – Bogalusa combined statistical area has a population of 1,360,436 as of 2000. The city/parish alone has a population of 343,829 as of 2010.

The city is named after Philippe d' Orléans, Duke of Orléans, Regent of France, and is well known for its distinct French Creole architecture, as well as its cross cultural and multilingual heritage.[2] New Orleans is also famous for its cuisine, music (particularly as the birthplace of jazz),[3][4] and its annual celebrations and festivals, most notably *Mardi Gras*. The city is often referred to as the "most unique"[5] city in America.[6][7][8][9][10]

New Orleans is located in southeastern Louisiana, straddling the Mississippi River. The boundaries of the city and **Orleans Parish** (French: *paroisse d'Orléans*) are coterminous.[11] The city and parish are bounded by the parishes of St. Tammany to the north, St. Bernard to the east, Plaquemines to the south and Jefferson to the south and west.[11][12][13] Lake Pontchartrain, part of which is included in the city limits, lies to the north and Lake Borgne lies to the east.[13]

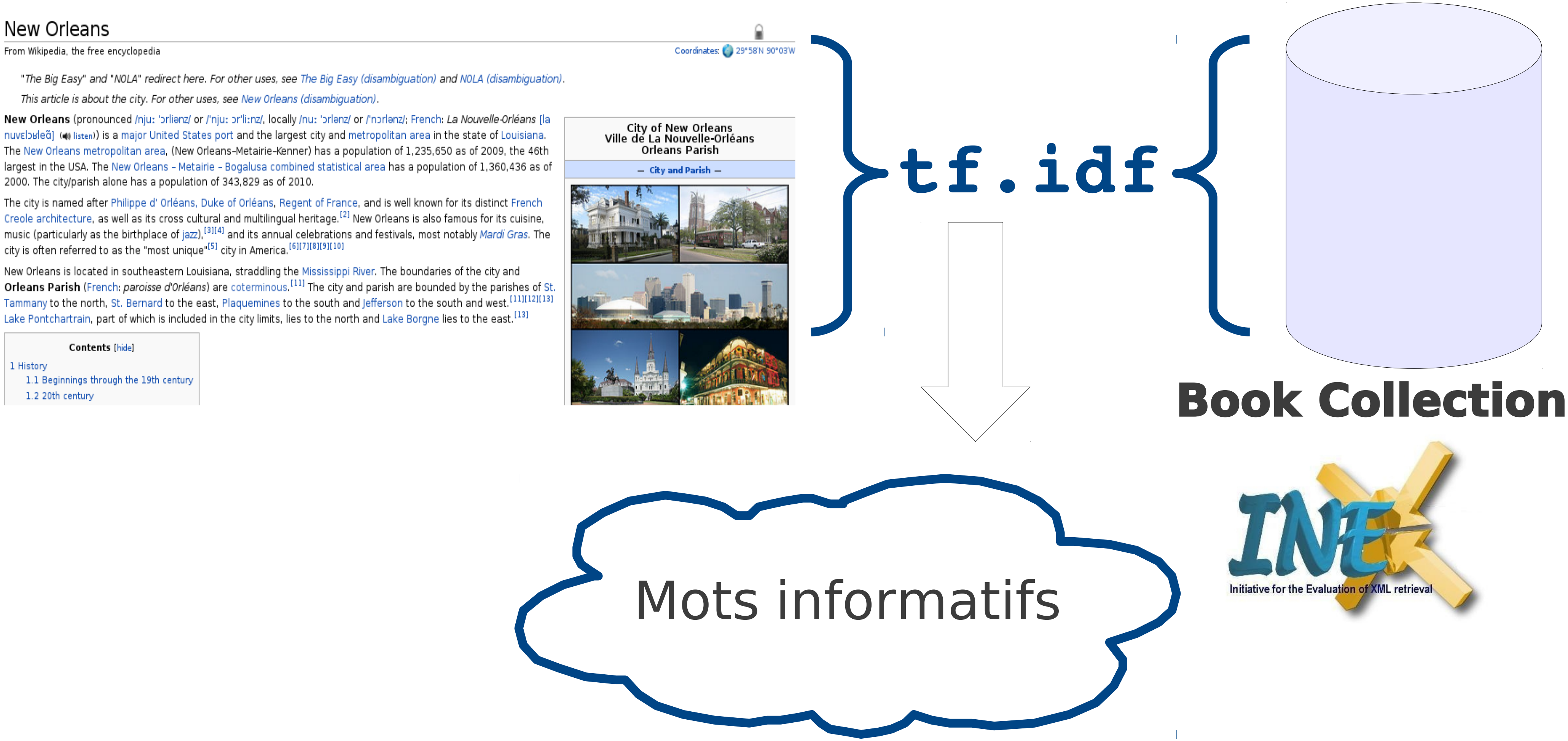
Contents [hide]

City of New Orleans
Ville de La Nouvelle-Orléans
Orleans Parish

— City and Parish —

Enrichissement de requêtes avec Wikipédia

Koolen *et al.* (WSDM'09)



New Orleans

From Wikipedia, the free encyclopedia

Coordinates: 29°58′N 90°03′W

"The Big Easy" and "NOLA" redirect here. For other uses, see The Big Easy (disambiguation) and NOLA (disambiguation).

This article is about the city. For other uses, see New Orleans (disambiguation).

New Orleans (pronounced /njuː ˈɔrliənz/ or /ˈnjuː ɔrˈliːnz/, locally /nuː ˈɔrlənz/ or /ˈnɔrlənz/; French: *La Nouvelle-Orléans* [la nuvɛlɔʁleɑ̃] (listen)) is a major United States port and the largest city and metropolitan area in the state of Louisiana. The New Orleans metropolitan area, (New Orleans–Metairie–Kenner) has a population of 1,235,650 as of 2009, the 46th largest in the USA. The New Orleans – Metairie – Bogalusa combined statistical area has a population of 1,360,436 as of 2000. The city/parish alone has a population of 343,829 as of 2010.

The city is named after Philippe d' Orléans, Duke of Orléans, Regent of France, and is well known for its distinct French Creole architecture, as well as its cross cultural and multilingual heritage.[2] New Orleans is also famous for its cuisine, music (particularly as the birthplace of jazz),[3][4] and its annual celebrations and festivals, most notably *Mardi Gras*. The city is often referred to as the "most unique"[5] city in America.[6][7][8][9][10]

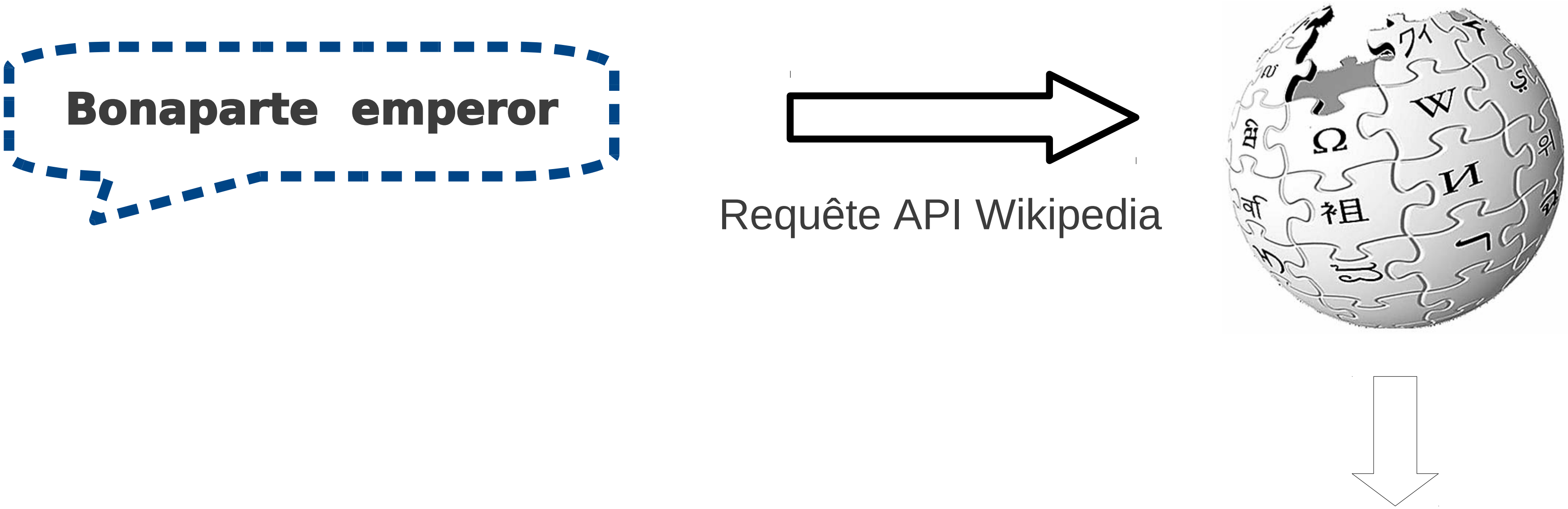
New Orleans is located in southeastern Louisiana, straddling the Mississippi River. The boundaries of the city and **Orleans Parish** (French: *paroisse d'Orléans*) are coterminous.[11] The city and parish are bounded by the parishes of St. Tammany to the north, St. Bernard to the east, Plaquemines to the south and Jefferson to the south and west.[11][12][13] Lake Pontchartrain, part of which is included in the city limits, lies to the north and Lake Borgne lies to the east.[13]

City of New Orleans
Ville de La Nouvelle-Orléans
Orleans Parish

— City and Parish —

Contents [hide]

Enrichissement de requêtes avec Wikipédia

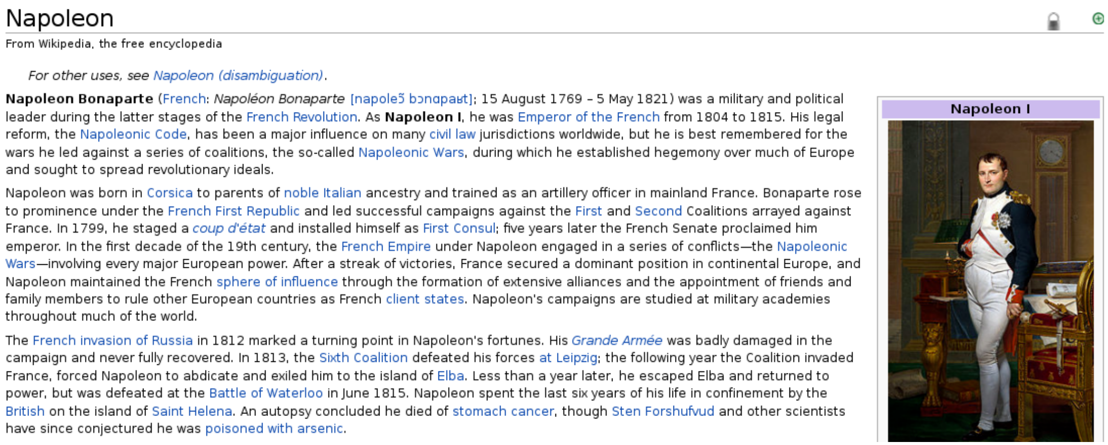


Bonaparte emperor

Requête API Wikipedia

`http://en.wikipedia.org/wiki/Napoleon`

Enrichissement de requêtes avec Wikipédia



Napoleon

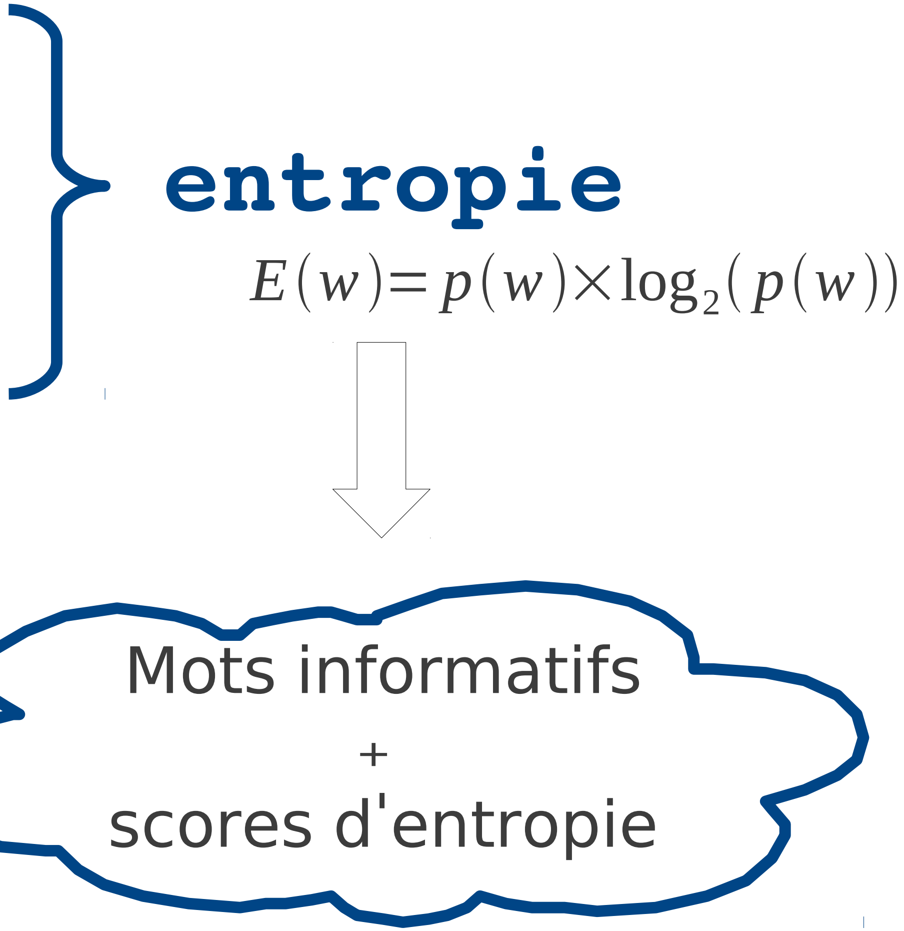
From Wikipedia, the free encyclopedia

For other uses, see Napoleon (disambiguation).

Napoleon Bonaparte (French: *Napoléon Bonaparte* [napoleɔ̃ bɔnapaʁt]; 15 August 1769 – 5 May 1821) was a military and political leader during the latter stages of the French Revolution. As **Napoleon I**, he was Emperor of the French from 1804 to 1815. His legal reform, the Napoleonic Code, has been a major influence on many civil law jurisdictions worldwide, but he is best remembered for the wars he led against a series of coalitions, the so-called Napoleonic Wars, during which he established hegemony over much of Europe and sought to spread revolutionary ideals.

Napoleon was born in Corsica to parents of noble Italian ancestry and trained as an artillery officer in mainland France. Bonaparte rose to prominence under the French First Republic and led successful campaigns against the First and Second Coalitions arrayed against France. In 1799, he staged a *coup d'état* and installed himself as First Consul; five years later the French Senate proclaimed him emperor. In the first decade of the 19th century, the French Empire under Napoleon engaged in a series of conflicts—the Napoleonic Wars—involving every major European power. After a streak of victories, France secured a dominant position in continental Europe, and Napoleon maintained the French sphere of influence through the formation of extensive alliances and the appointment of friends and family members to rule other European countries as French client states. Napoleon's campaigns are studied at military academies throughout much of the world.

The French invasion of Russia in 1812 marked a turning point in Napoleon's fortunes. His *Grande Armée* was badly damaged in the campaign and never fully recovered. In 1813, the Sixth Coalition defeated his forces at Leipzig; the following year the Coalition invaded France, forced Napoleon to abdicate and exiled him to the island of Elba. Less than a year later, he escaped Elba and returned to power, but was defeated at the Battle of Waterloo in June 1815. Napoleon spent the last six years of his life in confinement by the British on the island of Saint Helena. An autopsy concluded he died of stomach cancer, though Sten Forshufvud and other scientists have since conjectured he was poisoned with arsenic.



entropie

$$E(w) = p(w) \times \log_2(p(w))$$

Mots informatifs

+

scores d'entropie

Enrichissement de requêtes avec Wikipédia

Répartition des poids entre la requête originale et son enrichissement

- X <requête originale> + Y <mots issus de Wikipédia>

Pondération des mots extraits de la page Wikipédia

- Certains mots sont plus importants que d'autres
- Utilisation du score d'entropie

Variations du nombre de mots extraits de la page Wikipedia

- N : nombre de mots utilisés dans l'enrichissement

Protocole expérimental

Indri[1] pour l'indexation de la collection de livres et l'extraction de documents

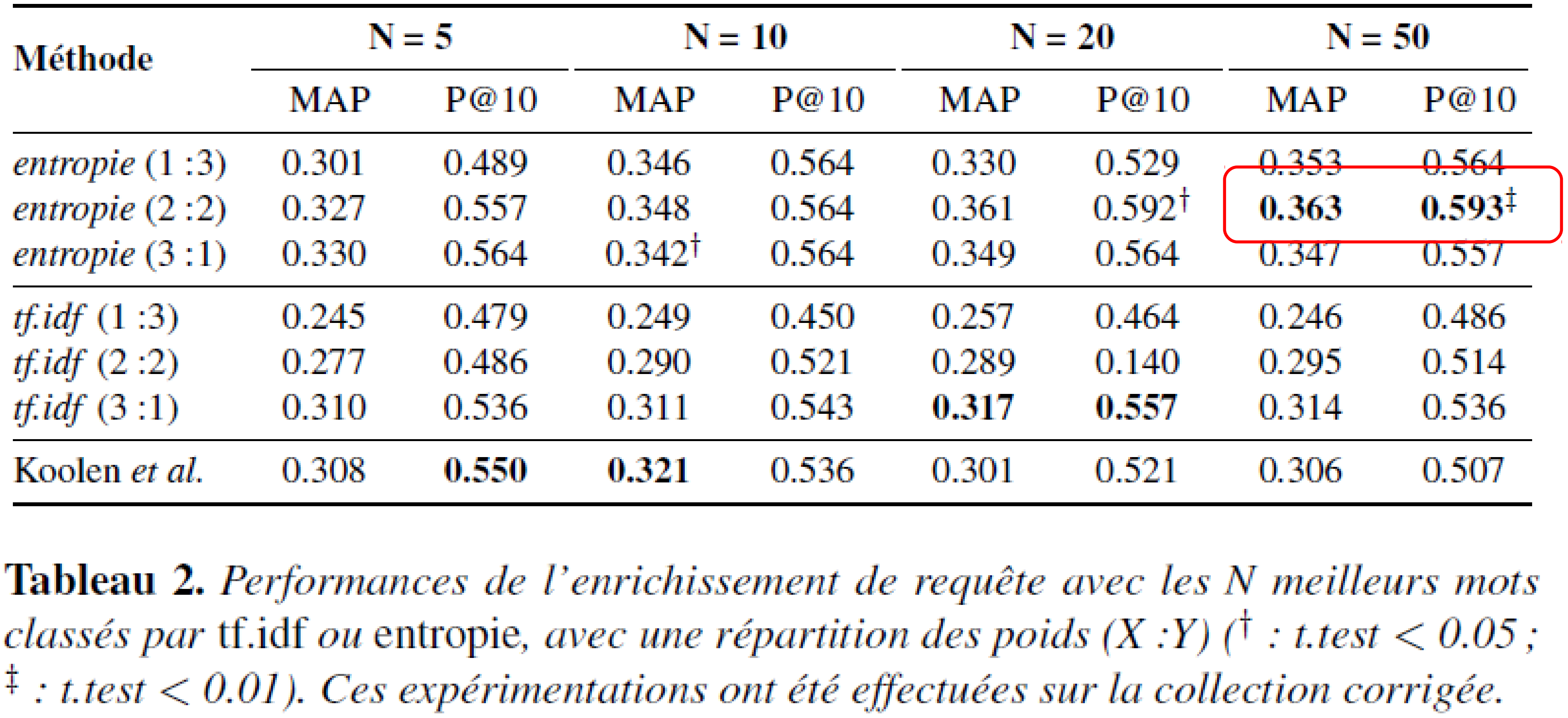
Développement d'une API : mirimiri[2]

- Sélection des pages Wikipédia ”en ligne”
- Calcul des scores d'*entropie*, *tf.idf* ...

[1] `http://www.lemurprojet.org`

[2] `http://mirimiri.org`

Résultats



Méthode	**N = 5**		**N = 10**		**N = 20**		**N = 50**	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
entropie (1 :3)	0.301	0.489	0.346	0.564	0.330	0.529	0.353	0.564
entropie (2 :2)	0.327	0.557	0.348	0.564	0.361	0.592†	**0.363**	**0.593**‡
entropie (3 :1)	0.330	0.564	0.342†	0.564	0.349	0.564	0.347	0.557
tf.idf (1 :3)	0.245	0.479	0.249	0.450	0.257	0.464	0.246	0.486
tf.idf (2 :2)	0.277	0.486	0.290	0.521	0.289	0.140	0.295	0.514
tf.idf (3 :1)	0.310	0.536	0.311	0.543	**0.317**	**0.557**	0.314	0.536
Koolen *et al.*	0.308	**0.550**	**0.321**	0.536	0.301	0.521	0.306	0.507

Tableau 2. *Performances de l'enrichissement de requête avec les N meilleurs mots classés par* tf.idf *ou* entropie, *avec une répartition des poids (X :Y) († : t.test < 0.05 ; ‡ : t.test < 0.01). Ces expérimentations ont été effectuées sur la collection corrigée.*

Conclusion

- Méthode de correction de césures dans les livres
  - Peu coûteuse
  - Amélioration des résultats de recherche de livres
- Enrichissement de requêtes
  - Wikipédia comme base de connaissances externe
  - Meilleurs résultats avec *entropie* et sans pondération entre la requête originale et l'enrichissement
- Extension de l'approche à d'autres types de documents : tâche Ad Hoc d'INEX

Merci de votre attention

Enrichissement de requêtes avec Wikipédia

- Répartition des poids entre la requête originale et son enrichissement
  - X : poids de la requête
  - Y : poids de l'enrichissement
- Pondération des mots extraits de la page Wikipédia
  - Certains mots sont plus importants que d'autres
  - Utilisation du score d'entropie
- Variations du nombre de mots extraits de la page Wikipedia
  - N : nombre de mots utilisés dans l'enrichissement

$$\Delta_Q(D) = \left(\prod_{i=1}^{k} p_D(q_i)^{\frac{1}{k}} \right)^{\frac{X}{X+Y}} \times \left(\prod_{i=1}^{n} p_D(t_i)^{\frac{w_i}{\sum_{j=1}^{n} w_j}} \right)^{\frac{Y}{X+Y}}$$