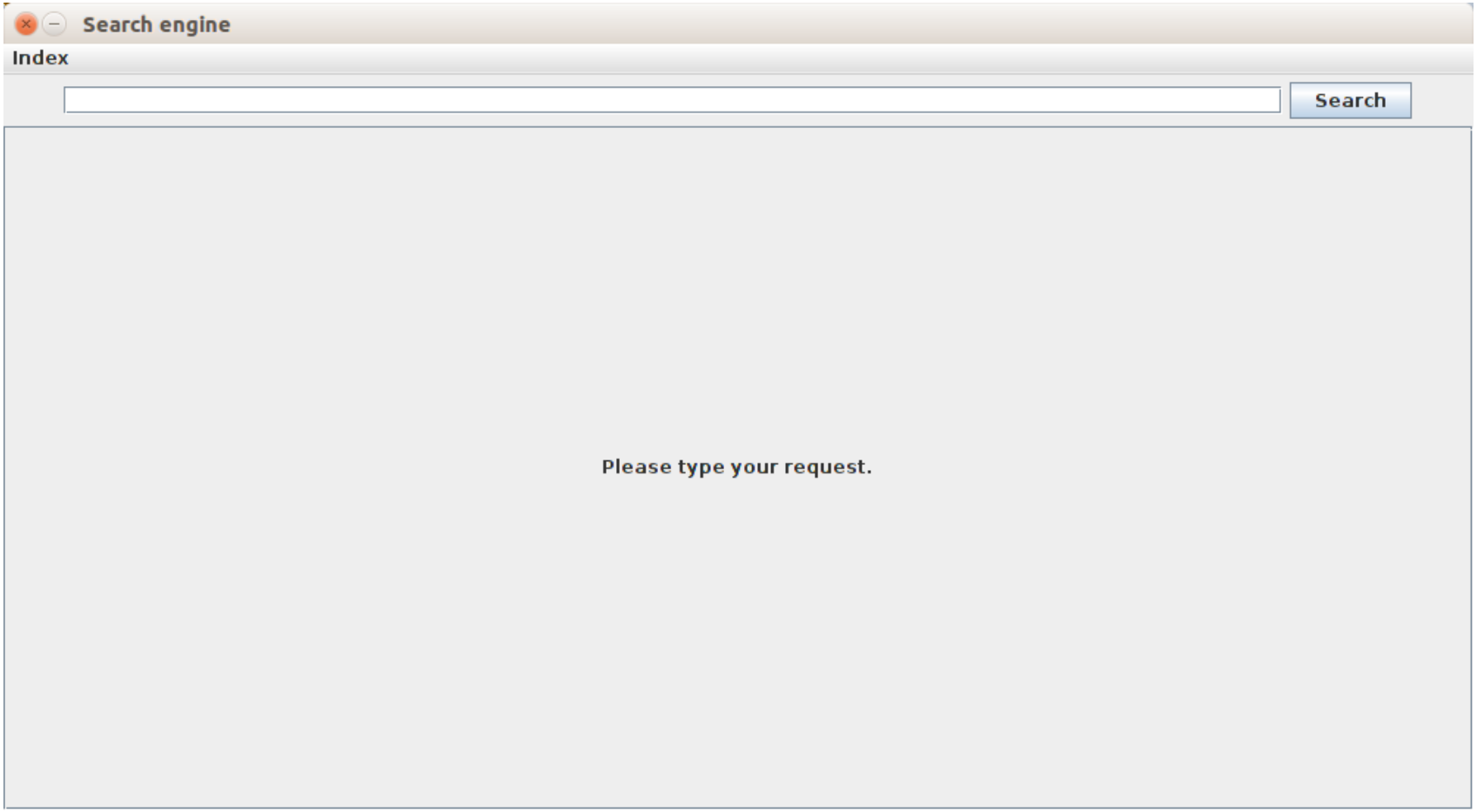
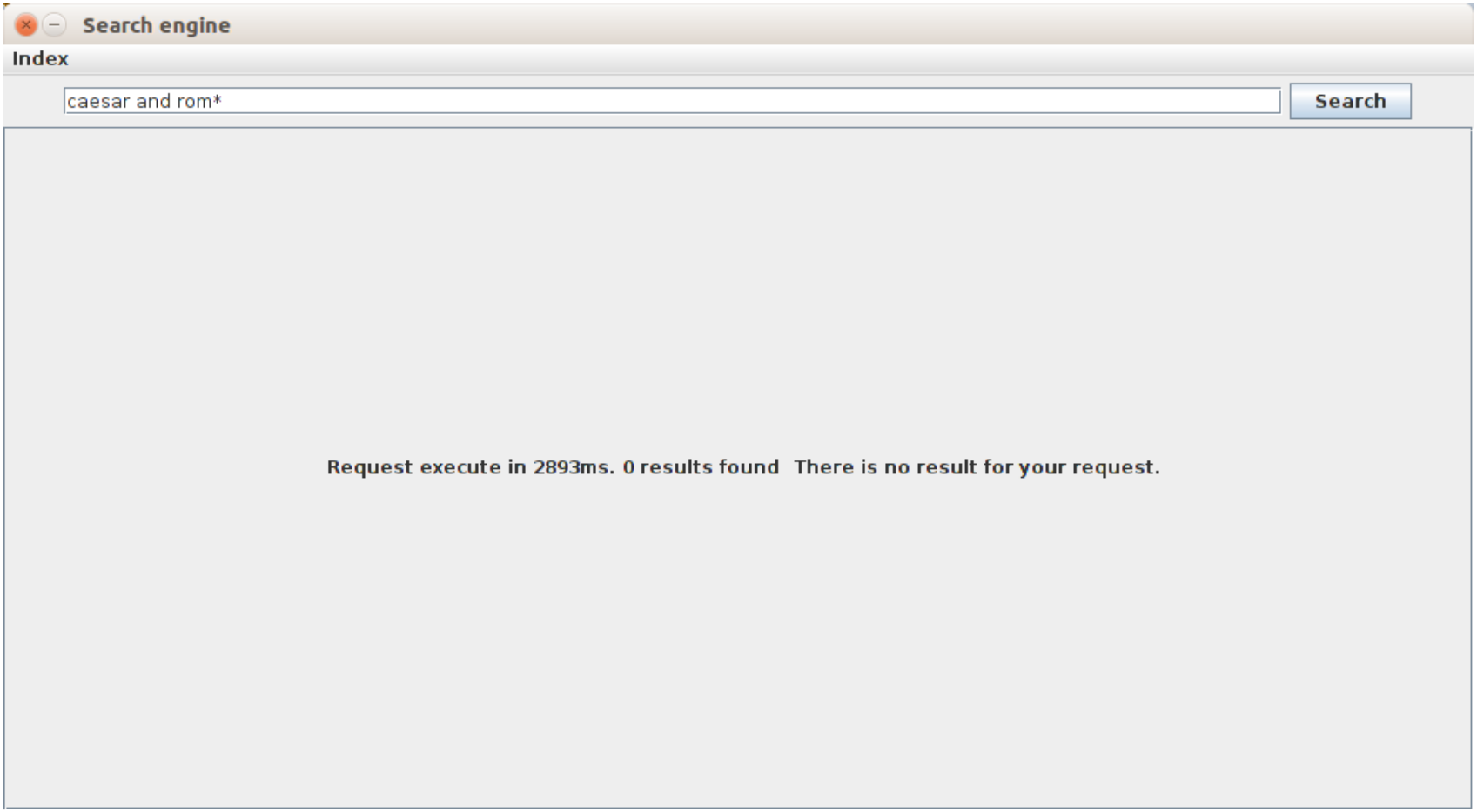
Recherche Documentaire

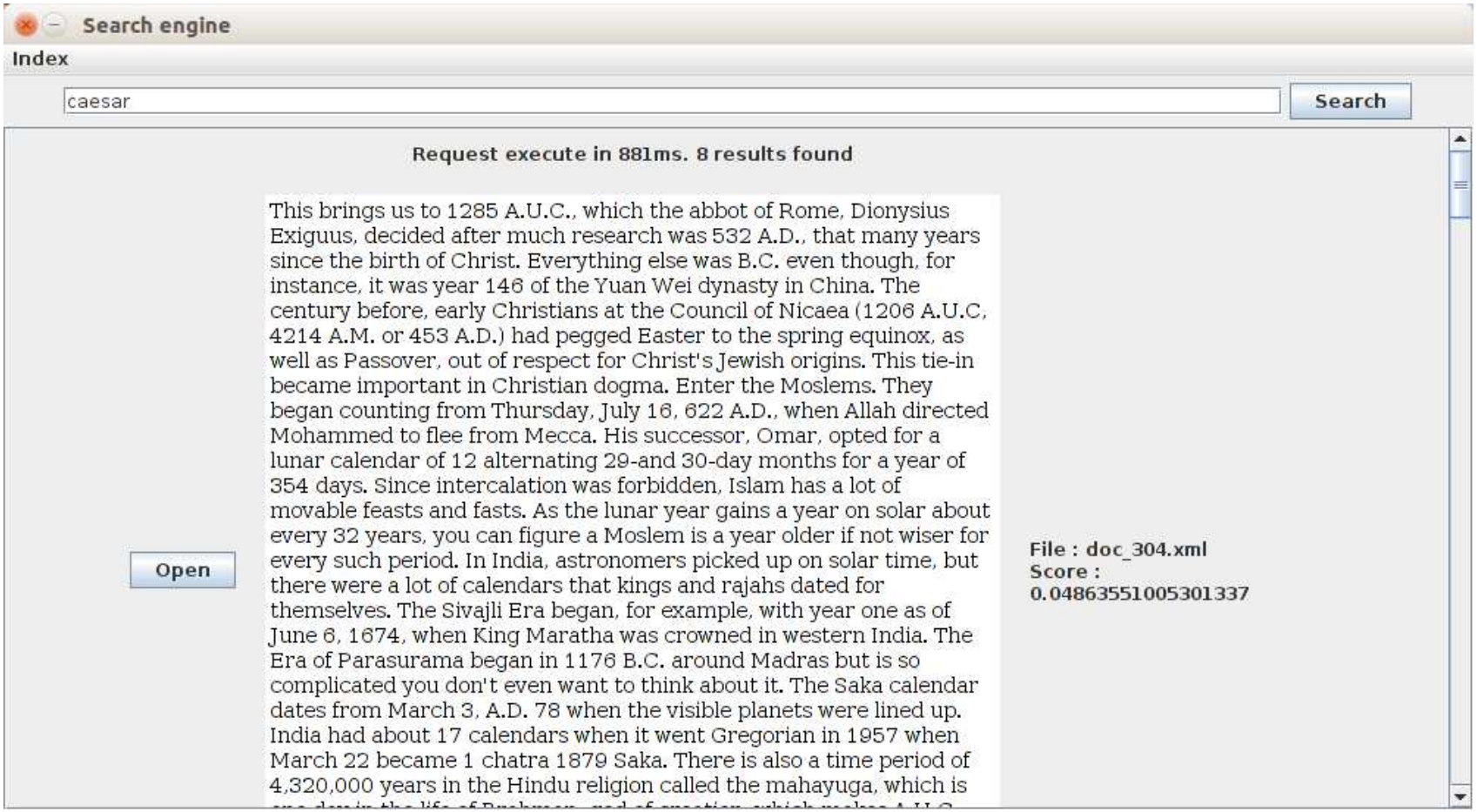
Jean-Baptiste Dalle & Romain Gaborieau

23 mars 2015

Introduction




Search engine
Index
caesar and rom*
Search
Request execute in 2893ms. 0 results found There is no result for your request.


Search engine
Index
caesar
Search
Request execute in 881ms. 8 results found
This brings us to 1285 A.U.C., which the abbot of Rome, Dionysius Exiguus, decided after much research was 532 A.D., that many years since the birth of Christ. Everything else was B.C. even though, for instance, it was year 146 of the Yuan Wei dynasty in China. The century before, early Christians at the Council of Nicaea (1206 A.U.C, 4214 A.M. or 453 A.D.) had pegged Easter to the spring equinox, as well as Passover, out of respect for Christ's Jewish origins. This tie-in became important in Christian dogma. Enter the Moslems. They began counting from Thursday, July 16, 622 A.D., when Allah directed Mohammed to flee from Mecca. His successor, Omar, opted for a lunar calendar of 12 alternating 29-and 30-day months for a year of 354 days. Since intercalation was forbidden, Islam has a lot of movable feasts and fasts. As the lunar year gains a year on solar about every 32 years, you can figure a Moslem is a year older if not wiser for every such period. In India, astronomers picked up on solar time, but there were a lot of calendars that kings and rajahs dated for themselves. The Sivajli Era began, for example, with year one as of June 6, 1674, when King Maratha was crowned in western India. The Era of Parasurama began in 1176 B.C. around Madras but is so complicated you don't even want to think about it. The Saka calendar dates from March 3, A.D. 78 when the visible planets were lined up. India had about 17 calendars when it went Gregorian in 1957 when March 22 became 1 chatra 1879 Saka. There is also a time period of 4,320,000 years in the Hindu religion called the mahayuga, which is
Open
File : doc_304.xml
Score :
0.04863551005301337

Corpus

- Un document par fichier
- Document au format XML
- Prise en compte uniquement de la partie "corps du texte"

- Plusieurs modèles possibles
- Choix du modèle vectoriel

"caesar" $\rightarrow$ D1 $\{3, 56\} \rightarrow$ D3 $\{7\}$

"world" $\rightarrow$ D6 $\{1, 5, 6\} \rightarrow$ D3 $\{8\} \rightarrow$ D8 $\{4\}$

- Suppression des caractères spéciaux
- Suppression des mots non spécifiques
- Réduction des mots à leur racine

Performances

L'indexation peut prendre un certain temps. Dans le cadre du projet, il n'est lancé qu'au premier démarrage de l'application ou sur demande.

Exemple

caesar is the king *and* dog *not* world

- Une requête est considérée comme un document à part entière
- Découpage en sous-requêtes
- Le même procédé de racinisation et de suppression des stopwords est réalisé sur chaque sous-requête

Exemple

caesar is the king *and* dog *not* world

Découpage

caesar ~~is the~~ king
and dog
not world

Exemple

caesar is the king *not* rom*

- L'indexation associe toutes les variantes non stemmées à leur racine
- Utilise les *regex* pour matcher le pattern
- *rom** peut correspondre à *roma* comme à *roman*

- Calcul du TF d'un mot dans un document : $\log_{10}(\frac{nbOccurences}{nbMot})$
- Calcul de l'IDF d'un mot dans le corpus : $\log_{10}(\frac{tailleDuCorpus}{nbOccurences})$
- Le TF-IDF d'un mot dans un document est alors calculé à partir de ces données

- Représentation des documents sous forme de vecteur
- Calcul de la proximité grâce au cosinus de Salton
- Plus la réponse est proche de 0, plus le document est pertinent

Les extensions

- Une correction orthographique est proposée
- Affichage dans la console
- Amélioration en sélectionnant automatiquement une correction ?

Conclusion

Avez-vous des questions ?