ПРОГРАММИРОВАНИЕ CUDA C/C++, АНАЛИЗ ИЗОБРАЖЕНИЙ И DEEP LEARNING

Лекция №5

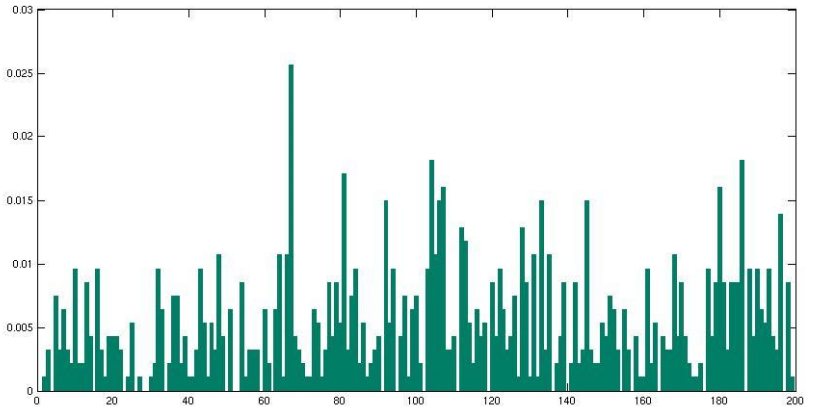
Спасёнов Алексей

Введение анализ текста

Модель bag-of-words (мешок слов)

Извлечение признаков

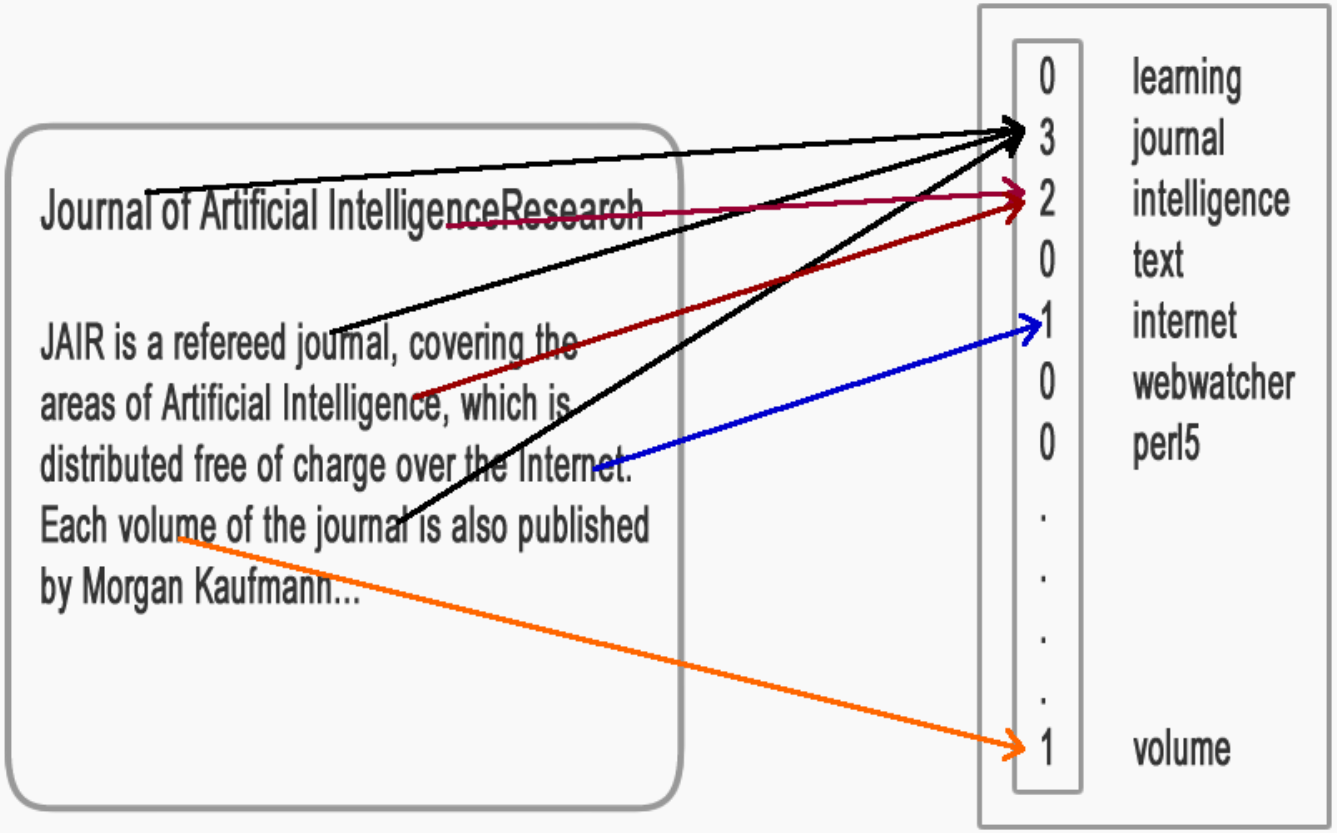
Bag-of-words



Анализ

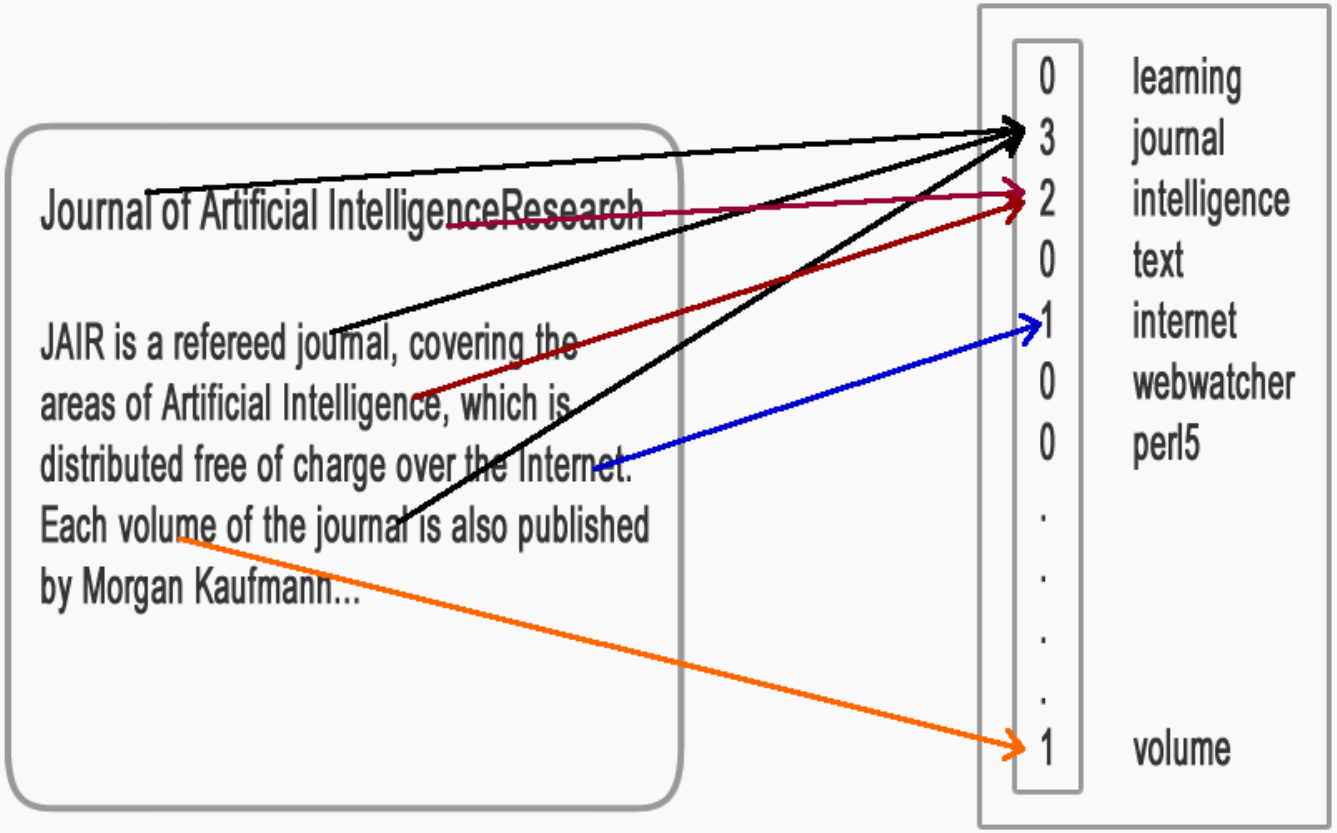
Введение анализ текста

Модель bag-of-words (мешок слов)



Введение анализ текста

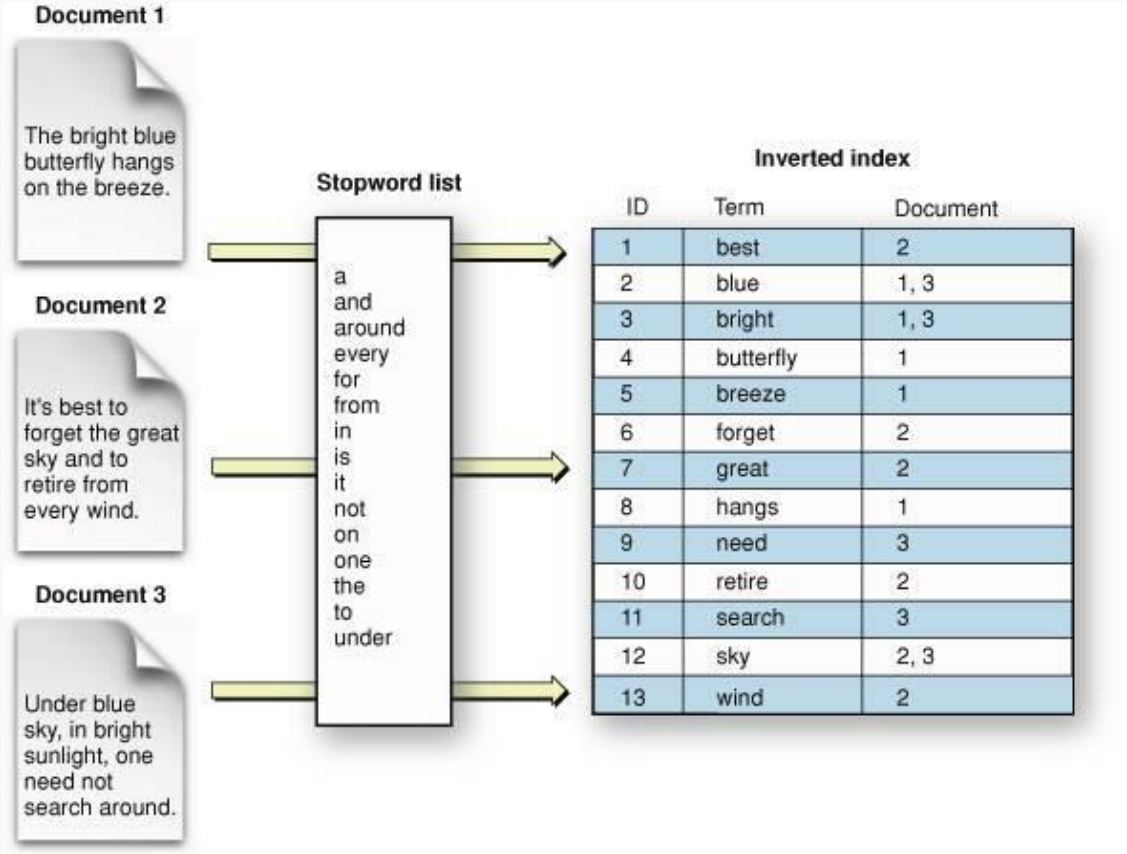
Модель bag-of-words (мешок слов)



Какая может быть проблема?

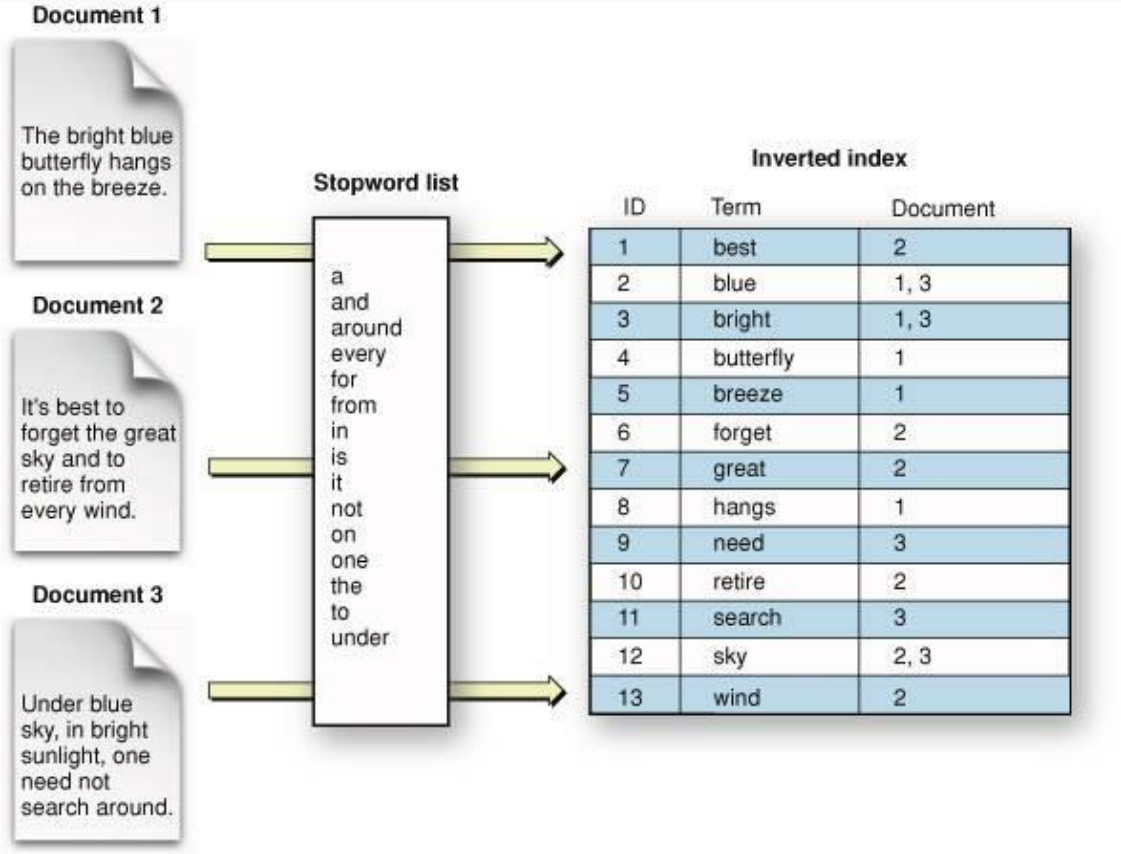
Введение анализ текста

Модель bag-of-words (мешок слов)



Введение анализ текста

Модель bag-of-words (мешок слов)

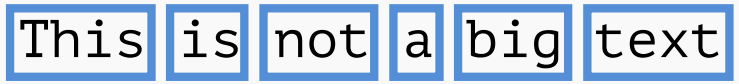


Но где же связь между словами?

Введение анализ текста

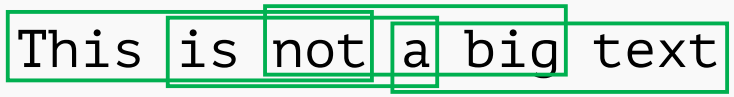
N-граммы

Unigrams из слов:



Bigrams из слов:

Trigrams из слов:



Введение анализ текста

N-граммы из букв

Unigrams:

text

t	2
e	1
x	1

Bigrams:

text

te	1
ex	1
xt	1

Trigrams:

text

tex	1
ext	1

Введение анализ текста

TF-IDF

TF (term frequency – частота слова)

Отношение числа вхождений некоторого слова к общему числу слов документа.

$$tf_{t,d} = \frac{n_t}{\sum_d n_d},$$

n_t - число вхождений слова t в документ

$\sum_d n_d$ - общее количество слов в документе

Введение анализ текста

TF-IDF

IDF (inverse document frequency – обратная частота документа)

Инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$idf_{t,D} = \log(\frac{|D|}{|\{d_i \in D | t_i \in d_i\}|})$$

$|D|$-количество документов в корпусе

$|\{d_i \in D | t_i \in d_i\}|$ - число документов в коллекции, в которых встречается слово t_i

$$tf - idf_{t,d,D} = tf_{t,d} * idf_{t,D}$$

Контакты:
a.spasenov@corp.mail.ru
alex_spasenov (Skype)

Спасибо за внимание!