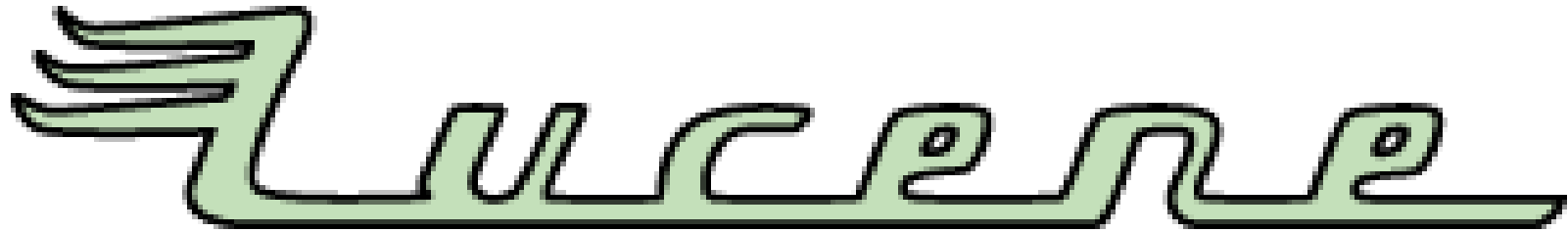


Context Vector & Document Preprocessing using Apache Lucene



Martin Kováčik

mato.kovacik@gmail.com

<http://student.fiit.stuba.sk/~kovacik03/dipl.html>

Context Vector

- ◆ Význam slova je určený lingvistickým kontextom
- ◆ Slová, ktoré sú príbuzné sa vyskytujú v rovnakých kontextoch
- ◆ Podobnosť slov je možné určiť na základe ich spoločného výskytu v rovnakých kontextoch

Context Vector

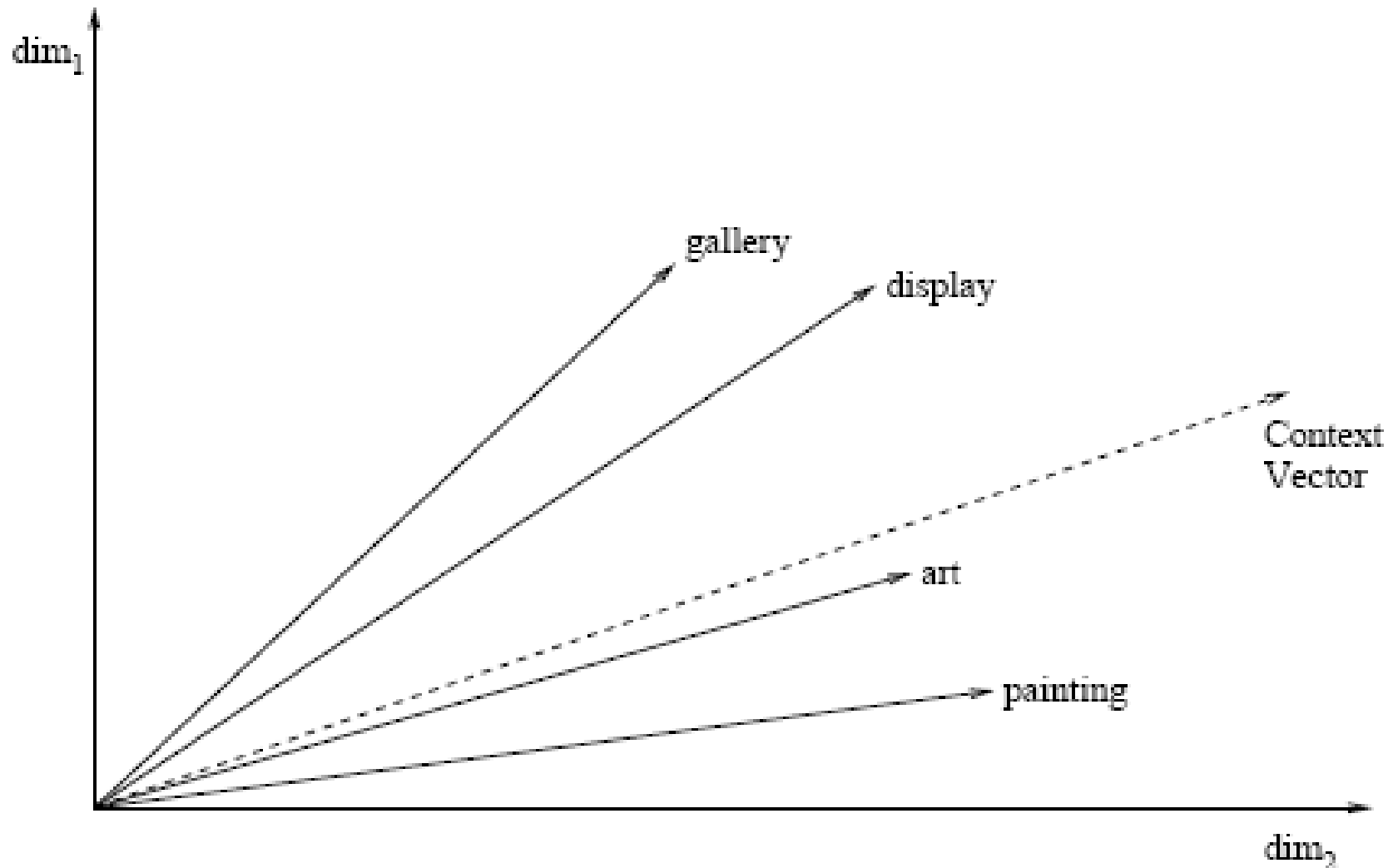
- ◆ Vzťah medzi slovami je možné vyjadriť ako počet spoločných výskytov v kontextoch
- ◆ Vzťahy a ich ohodnotenia vytvárajú priestor
- ◆ Takýto priestor je pomerne riedky
- ◆ Je potrebné odstrániť nadbytočné slová tzv. STOP words – *druhá časť prezentácie*

Context Vector

- ◆ Slová, ktoré sú podobné vytvárajú zhluky v tomto priestore
- ◆ Je možné vyhodnocovať ich vzdialenosť ako mieru podobnosti
- ◆ Vytváranie vektorov pre kontext skladaním vektorov pre slová

Context Vector

The paintings were displayed in the art gallery



Context Vector

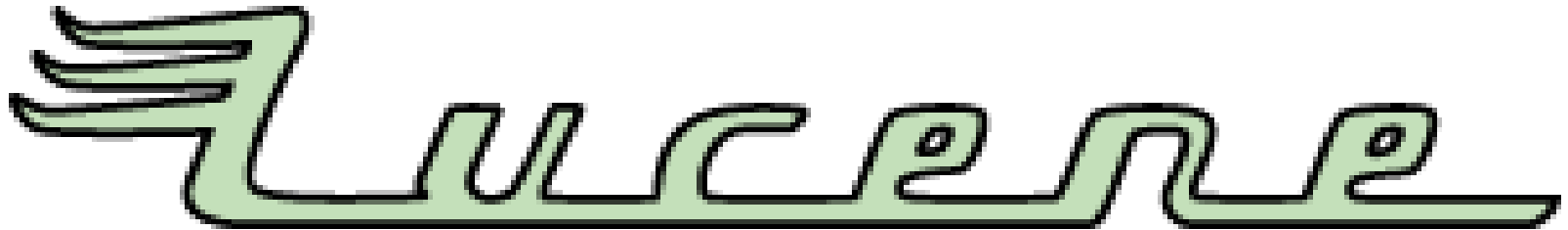
- ◆ Priestor slov je možné vytvárať z existujúcich dokumentov (popr. Databáz - WordNet)
- ◆ Univerzálna aplikácia – metóda nie je úzko prepojená s problémovou doménou porovnávaných dokumentov
- ◆ Porovnanie je možné realizovať meraním vzdialenosti. Nad vytvoreným priestorom je možné realizovať ďalšie operácie.

Document Preprocessing - Lucene

- ◆ Ciel':
 - ◆ Rozbitie textu na slová
 - ◆ Normalizácia slov
 - ◆ Spracovanie slova na základný tvar
 - ◆ Odstránenie nadbytočných slov
 - ◆ Nahradenie synonym
- ◆ Uvedené ciele sa prekrývajú s cieľmi pri indexácii dokumentov

Document Preprocessing - Lucene

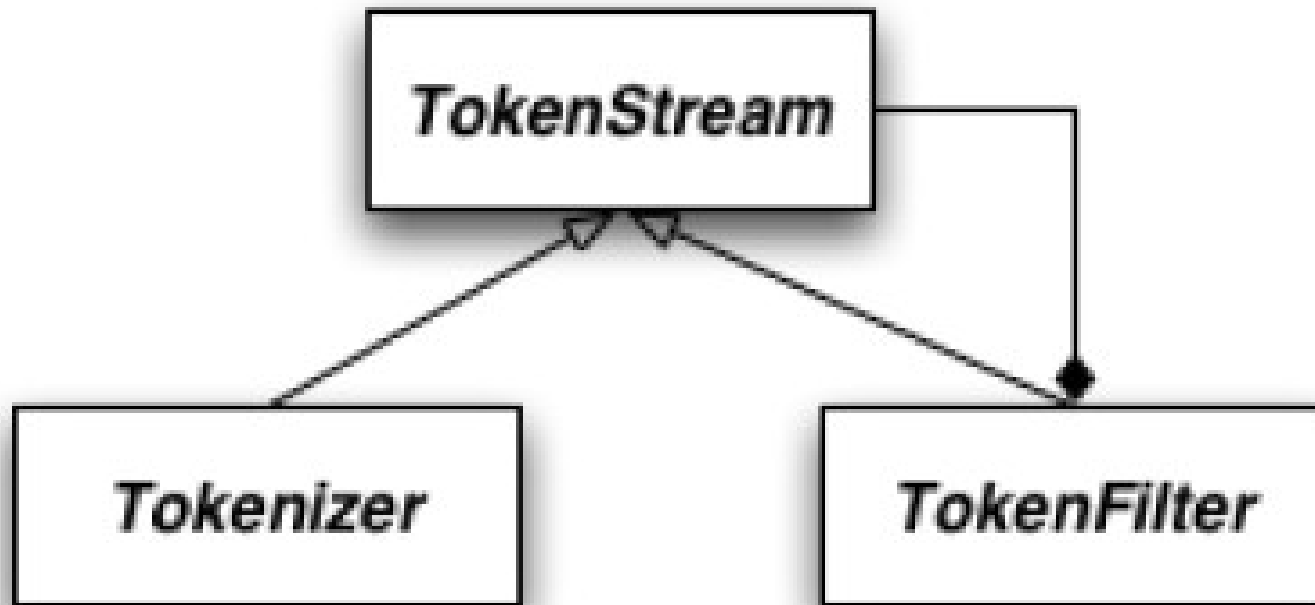
- ◆ Apache Lucene <http://lucene.apache.org>



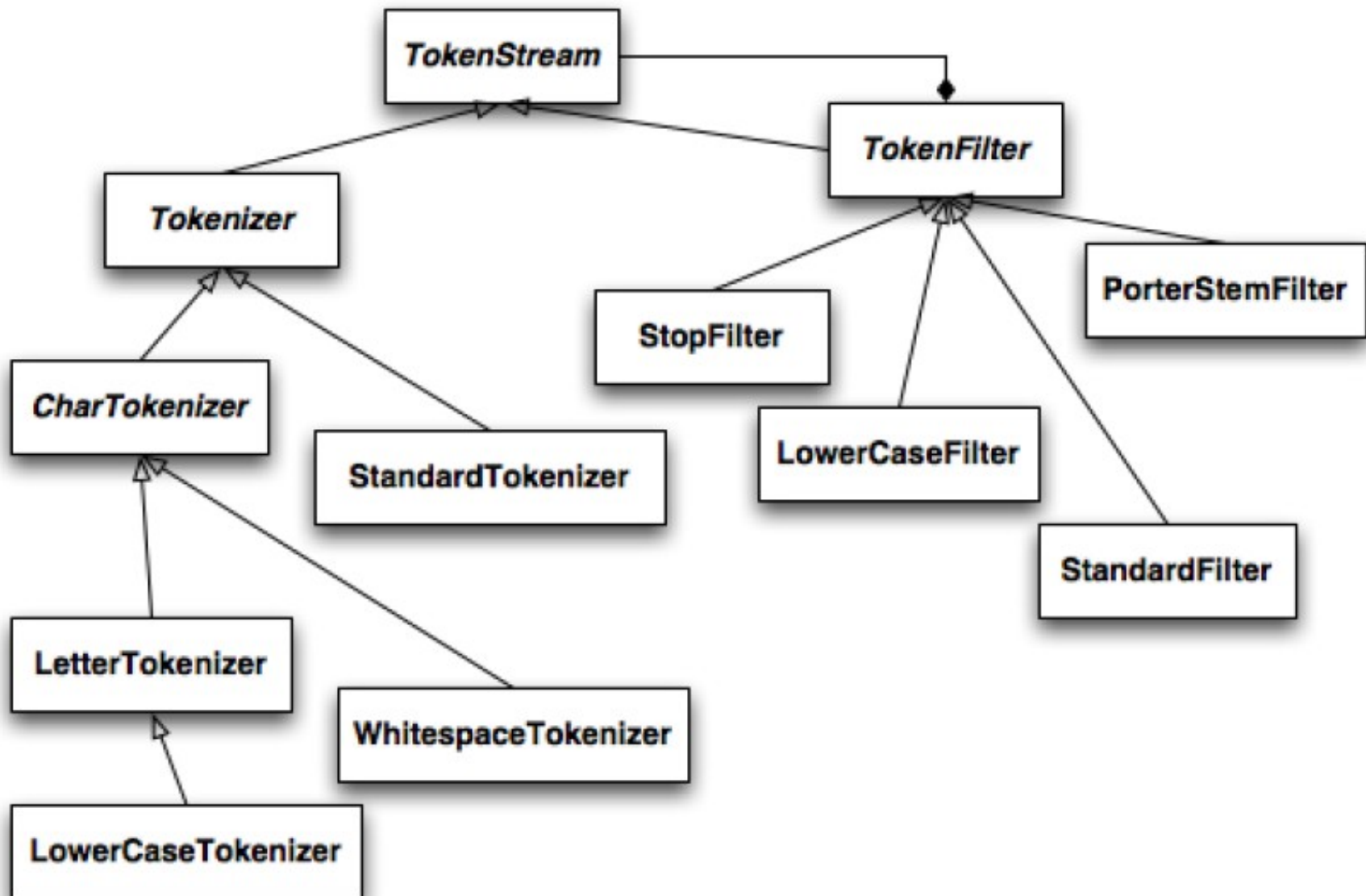
- ◆ Spracovanie textu prostredníctvom skladateľných dátovodov a filtrov

Document Preprocessing - Lucene

- ◆ TokenStream, Tokenizer & TokenFilter



Document Preprocessing - Lucene



Document Preprocessing - Lucene

Ciel':

Rozbitie textu na slová

Normalizácia slov

Spracovanie slova na
základný tvar

Odstránenie nadbytočných
slov

Riešenie:

StandartTokenizer

LowerCaseFilter

PorterStemFilter

StopFilter

Document Preprocessing - Lucene

- ◆ **Dalšie možnosti – Lucene Sandbox**
 - ◆ Snowball Stemmers for Lucene
 - ◆ Contributed Analyzers, Tokenizers, and Filters for various languages
 - ◆ WordNet/Synonyms
 - ◆ High Frequency Terms – automatické zostavovanie zoznamov STOP words

Referencie

- ◆ Patwardhan and Pedersen: Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts - EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, April 4, 2006, Trento, Italy.
- ◆ Lucene: <http://lucene.apache.org>
- ◆ Otis Gospodnetič, Erik Hatcher: Lucene in Action, Manning, 2005

Context Vector & Document preprocessing

Dakujem za pozornosť

mato.kovacik@gmail.com

<http://student.fiit.stuba.sk/~kovacic03/dipl.html>