Text Characteristics

Jan Rygl Aleš Horák

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

xrygl@fi.muni.cz, hales@fi.muni.cz

Workshop of
Seznam.cz & Natural Language Processing Centre
May 16, 2014

Outline

Keyword/Keyphrases/Topics extraction

Example (http://zpravy.ihned.cz/c1-59700380-smrtici-naloze-byly-v-bostonu-dve)

Smrtící nálože byly v Bostonu dvě.

Při útocích v Bostonu zemřeli tři lidé, dalších 150 exploze zranily. Americký prezident Barack Obama prohlásil, že šlo o teroristický čin. Podle FBI sice bezprostřední nebezpečí dalších útoků nehrozí, na letištích je přesto znát nervozita – kvůli ohlášení podezřelého objektu bylo evakuováno například letiště LaGuardia.
V pondělí odpoledne deset minut před 15. hodinou vybuchly v Bostonu během maratonu podél trati dvě nálože. Původně se mluvilo zhruba o stovce zraněných.

...

Extracted keywords:

- FBI
- letiště
- bomba
- Boston
- Obama
- nálož

Keyword/Keyphrases/Topics extraction

Po deštích se rozvodnily Olše a Bečva, na Moravě a ve Slezsku dál prší Po osmé hodině ráno klesla hladina Olše v Českém Těšíně z třetího na druhý povodňový stupeň. Podle hydrologů hrozí, že třetího povodňového stupně dosáhne Bečva ve Valašském Meziříčí a Rožnově pod Radhoštěm. Na Moravě dále prší a déšť má trvat celý den. Kvůli popadaným stromům na řadě míst nejezdí vlaky. Rozvodněná Olše v Českém Těšíně v pátek ráno v 8 hodin mostu Svobody směrem na... Dalších 11 fotografií v galerii Rozvodněná Olše v Českém Těšíně v pátek ráno v 8 hodin mostu Svobody směrem na polskou část Těšína . Bdělost, tedy první stupeň povodňové aktivity , platí podle webu Českého hydrometeorologického ústavu (ČHMÚ) na Lubině v Petřvaldu, na Ostravici v Ostravě a Frýdku-Místku, na Olši ve Věřovicích a Jablunkově i na Bečvě v Teplicích nad Bečvou. Nejhorší je situace na Olši v Českém Těšíně v Moravskoslezském kraji, kde během

...

Detected topics are:

polská část Těšína (44.55)
Horní Bečva (44.45)
hladina Olše (42.41)
Povodí Moravy (30.78)
první stupeň povodňové aktivity (26.25)
odstraňování následků nehody (25.93)
24 hod (25.66)

Keyword/Keyphrases/Topics extraction

Definition

Words/phrases used to characterise the contents of a document.

Method

Select words/phrases that appear with statistically unusual frequency in a text

Applications

- Terminology extraction
- Text classification (topic, spam)
- Search Engine Optimisation (SEO)
- Text filtering (job advertising, RSS)
- Text summarization

Communication Pattern Analysis

Text characteristics

Word count:	47.61
Verb/Word ratio:	48.58
Verb/Adjective ratio:	47.24
Conditionals count:	43.64
Occasion/Restriction ratio:	66.37
Solution/Opinion ratio:	41.5
Word/Occasion ratio:	44.27
Word/Restriction ratio:	59.45
Word/Solution ratio:	51.14
Word/Opinion ratio:	43.91
Simple/Difficult ratio:	55.54
Aim:	49.49
Original/Indefinite ratio:	50.98
Keen/Sloppy ratio:	47.82
Vocabulary richness:	49.78

reaguji na Váš inzerát zveřejněný na internetových stránkách www.neziskovky.cz, ve kterém hledáte pracovníka fundraisingu a PR. Po posouzení Vámi stanovených požadavků předpokládám, že mohu být vhodným kandidátem.

Mohu nabídnout své několikaleté zkušenosti s prací s médii. V oblasti PR mám mimo jiné zkušenosti s psaním a vydáváním tiskových zpráv, s vystupováním v rozhlase a televizi, analýzou mediálních dopadů a prezentací. Taktéž jsem vedl několik kampaní pro neziskovou organizaci Přátelé zvířat. Pořádal jsem několik přednášek o lidských právech a kultuře v zemích Tibetu, Barmě a Arménii. Byl jsem členem produkčního týmu filmového festivalu Jeden svět v Praze v roce 2004. V organizaci Přátele zvířat jsem se podílel na získávání finančních prostředků z grantů a od individuálních dárců. Do prosince 2005 jsem spravoval databázi členů a přispěvatelů organizace včetně komunikace s nimi. Z činnosti pro Přátele zvířat mám četné zkušenosti z jednání se státní správou, politiky a zástupci firem.

Plně si uvědomuji důležitost získávání finančních prostředků pro činnosti vykonávané sítí diakonií po celé České republice. Diakonii vnímám jako specifickou organizaci s dlouholetou tradicí, jež byla v České republice přerušena obdobím komunismu. Nepovažuji se za věřícího člověka, avšak mám úctu ke křesťanským hodnotám a vážím si činností, jež diakonie vykonávají. Činnosti, kterými se zabývají diakonie jsou mi velmi blízké. Vámi

Author's traits

Vocabulary analysis

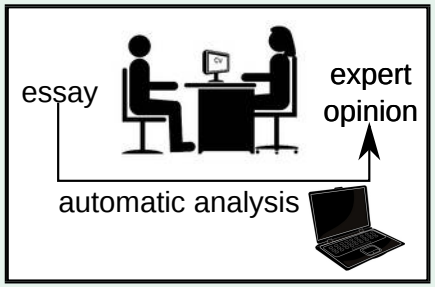
- expression of occasion `Nabízím` prezidentský úřad jako neutrální pole
- expression of solution Jediné reálné `řešení` je skutečně změnit politický systém
- relativizing words slyšeli `asi` všichni, kteří se alespoň `trochu` zajímají
- opinions/ideas abych `vyjádřil` uznání všem
- limits/constraints Pokud udělá více `restrikcí`, tak
- phrases Okamura nenechá `nit suchou`
- word repetition že `právě tito lidé` jsou solí země a že `právě tito lidé` jsou
- key words, unknown words
- modal verbs, conditionals

Communication Pattern Analysis

Motivation

- Analysis of personality traits using author's verbal style
- Optimize communication strategies
- Behaviour prediction

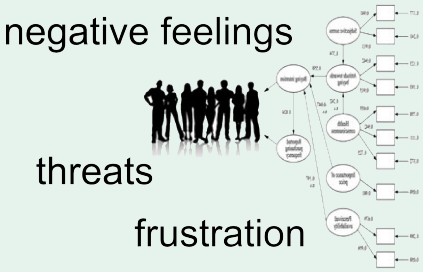
Applications



Job interviewing

a symbol of **power** and largest **flying** land bird to **travel** up to 250 km Condors are so **large** that

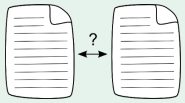
Brand naming



Safe workplace

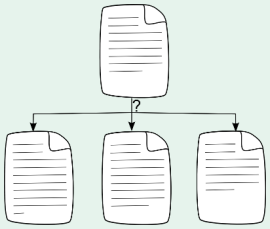
Problem Definition

Authorship Verification



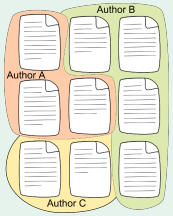
- decide if a document was written by the signed author

Authorship Attribution



- find out an author of a document
- candidate authors can be known

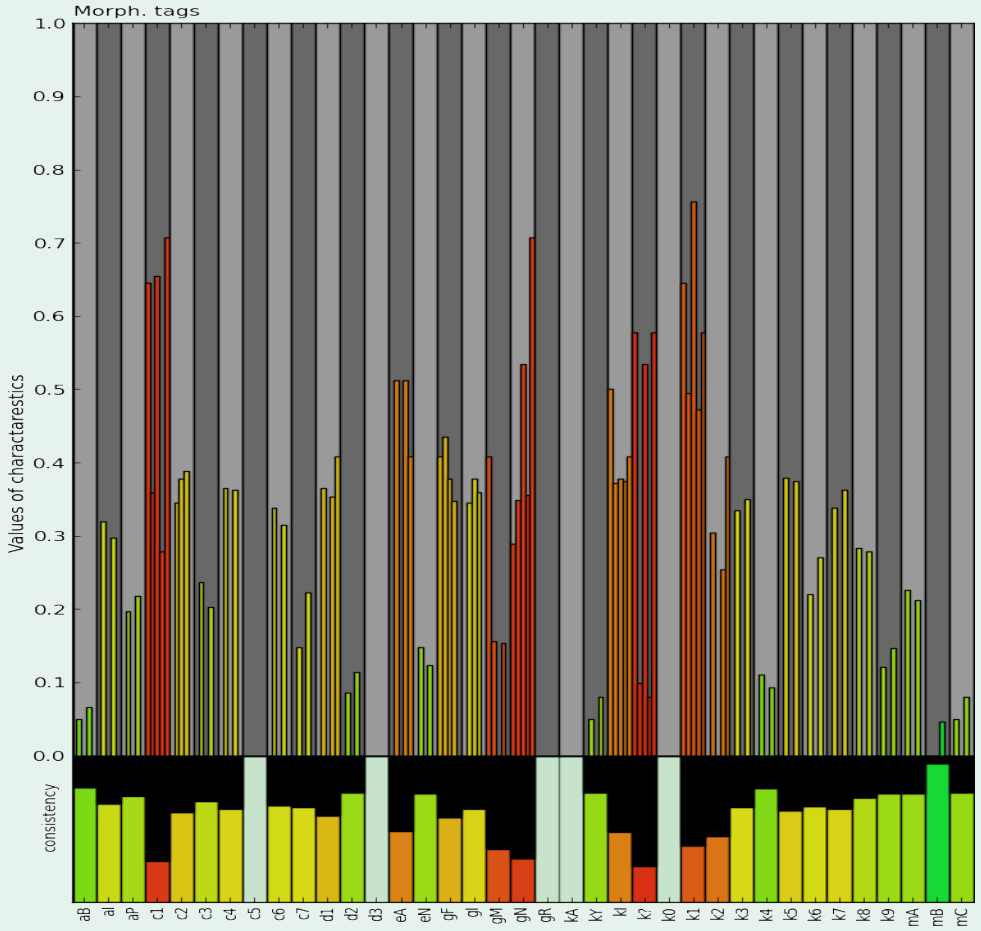
Authorship Clustering



- cluster documents or text paragraphs according to the authors

Author Writeprint/Stylom

Collection of author's documents

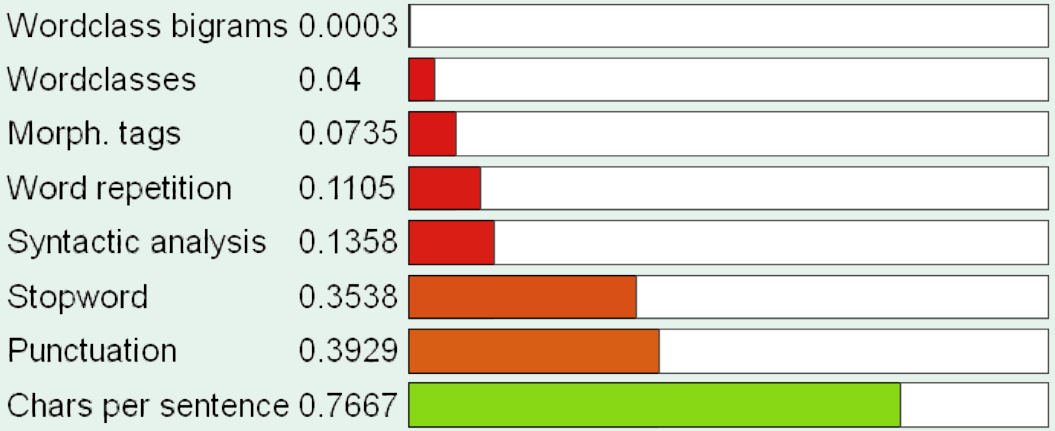


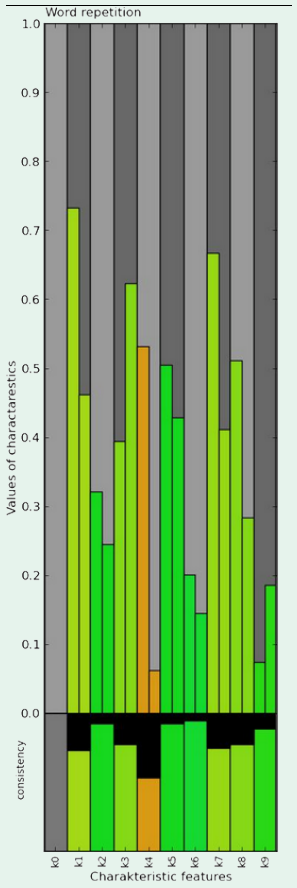
Author analysis:

1. Range: typical feature values for that author
2. Consistency (deviation): which features are most important

Authorship Verification

Stylometry





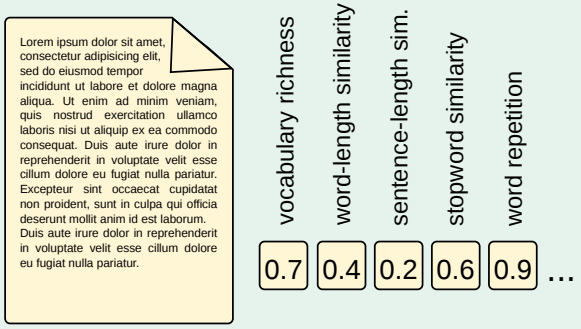
Example: comparison between two different authors (similarity: 4 %)

Other stylometric features: dialect, typography, errors, gender and age, vocabulary richness, . . .

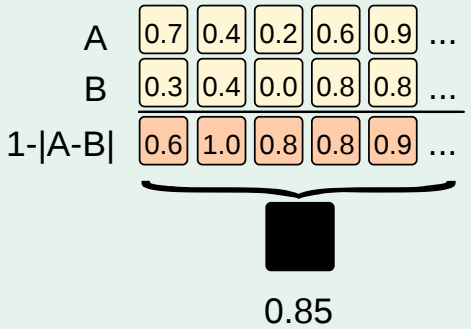
Machine learning approach

Similarity-based Machine Learning (verification)

1. Extract document features for each stylometric characteristic



2. Compare extracted features from documents to obtain similarities



3. Machine learning determines a probability of the same authorship

Accuracy

Verification:

- books, essays: 95 % – 100 %
- blogs, newspapers: 70 % – 90 %
- e-mails, sms: 60 % – 70 %

Attribution (depends on the number of candidates, comparison on blogs):

- up to 4 candidates: 85 % – 95 %
- up to 10 candidates: 65 % – 85 %
- up to 100 candidates: 40 % – 70 %

Conclusions

Keyword Extraction

A Brief representation of the content of a document.

Communication Pattern Analysis

An analysis of personality traits.

Authorship Recognition

An uncovering authorship of anonymous texts.