Data mining in the age of social web

Ákos Bardóczi

Downloads: https://bardoczi.net
Web: http://genetics.dote.hu/bardoczi
Email: bardoczi@med.unideb.hu
Tel: +1 202 4700-790
Fax: +36 1 9997-993
Comments: akos@cryptolab.net

If You Steal From One Author, It's Plagiarism;
If You Steal From Many, It's Research

unknown author

http://quoteinvestigator.com/2010/09/20/plagiarism/

Highlighted topics

- a part of open source intelligence
- simple data extraction/mining techniques from web 2.0 services
- authorship attribution/authorship profiling
- plagiarism detection

Old fashioned sources in nutshell

- WHOIS-records
- email long headers – which sent via web, commercial providers
- DNS-records
- patterns in cookies and HTTP-response streams

Old fashioned sources – WHOIS

- in case gTLD – WHOIS masking
- the fake contact details similars to each other between suspected domains - frequent
- private registrar companies owned by domain owner

Old fashioned sources
email long header

- The popular misbelief about Gmail, AOL, YahooMail an others

Old fashioned sources
email long header in fact

Return-path: <bardoczi@gmail.com>
Received: from st11p00mm-smtpin012.mac.com ([17.172.87.212]) by ms03531.mac.com (Oracle Communications Messaging Server 7u4-27.08 (7.0.4.27.7) 64bit (built Aug 22 2013)) with ESMTP id <0MUC007OQVDVDMG0@ms03531.mac.com> for bardoczi@me.com; Tue, 08 Oct 2013 15:08:19 +0000 (GMT)
Original-recipient: rfc822;bardoczi@me.com
Received: from mail-we0-f182.google.com ([74.125.82.182]) by st11p00mm-smtpin012.mac.com (Oracle Communications Messaging Server 7u4-27.08(7.0.4.27.7) 64bit (built Aug 22 2013)) with ESMTP id <0MUC00MUCVDUUTJ1@st11p00mm-smtpin012.mac.com> for bardoczi@me.com (ORCPT bardoczi@me.com); Tue, 08 Oct 2013 15:08:19 +0000 (GMT)
Received-SPF: pass (st11p00mm-smtpin001.mac.comdomain of bardoczi@gmail.com designates 74.125.82.182 as permitted sender) receiver=st11p00mm-smtpin001.mac.com; client-ip=74.125.82.182; helo=mail-we0-f182.google.com; envelope-from=bardoczi@gmail.com; x-software=spfmilter 0.97 http://www.acme.com/software/spfmilter/ with libspf-unknown;

the server which first recieved

Received: by mail-we0-f182.google.com with SMTP id t61so8978535wes.13 for <bardoczi@me.com>; Tue, 08 Oct 2013 08:08:17 -0700 (PDT)
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed; d=gmail.com; s=20120113; h=mime-version:reply-to:date:message-id:subject:from:to:content-type; bh=F7h8uU2A3blCvcj6XbYiBDxawl1FFa98TBwa/d1NMpM=;
b=zE5NNhuRYbJ1pF6FVzzV3sUlwYShaPj/b5aFpEAY4geEe0Avi+nYh45jZl66gaFZg/
FWm/WrVXOypzB6BDklqvQzXnYSq8XbP6BlJY4W4calkByv0Gi6RS5RSECV+goLCul4DT
RHfcKyKNCBoNSSL1lQdjjZaRNL/BrK0oNm5rhChSiepYurr4uqnkeUiHk5hpr6W4+Et8
ZC3J1vpa7I4Ln4fWvlG6p63olwZFMMmvghOuH3lenzRPZCl636z5PvvXYLorwWJc7gD
khshPztljkimUtWclm3Wgc50we+VEre3QKy+4rd8bEhSTlrXzJ2MUJtef27M3cbfZZPr pWYA==
MIME-version: 1.0
X-Received: by 10.180.93.166 with SMTP id cv6mr2026676wib.37.1381244897844; Tue, 08 Oct 2013 08:08:17 -0700 (PDT)

sender's timezone

Received: by 10.... with HTTP; Tue, 08 Oct 2013 08:08:17 -0700 (PDT)
Reply-to: akos@secure.bardoczi.net
Date: Tue, 08 Oct 2013 17:08:17 +0200
Message-id: <CAFMsfx1G4uXMG-B_TP0ihvYBhMRe3Zc8diQxbkoRXEtDM4TSEw@mail.gmail.com>
Subject: Nem mondom meg, hogy hol vagyok
From: =?ISO-8859-1?Q?=E1kos_bard=F3czi?= <bardoczi@gmail.com>
To: bardoczi@me.com
Content-type: multipart/alternative; boundary=f46d043c7fbc7befc104e83c23ec
Authentication-results: st11p00mm-smtpin012.mac.com; dkim=pass reason="2048-bit key" header.d=gmail.com header.i=@gmail.com header.b=zE5NNhuR; dkim-adsp=pass

character encoding – may indicates the senders language and nationality

x-icloud-spam-score: 33322 f=gmail.com;e=gmail.com;pp=ham;spf=pass;dkim=pass;wl=absent;pwl=absent
X-Proofpoint-Virus-Version: vendor=fsecure engine=2.50.10432:5.10.8794,1.0.431,0.0.0000 definitions=2013-10-08_06:2013-10-08,2013-10-08,1970-01-01 signatures=0
X-Proofpoint-Spam-Details: rule=notspam policy=default score=1 spamscore=1 suspectscore=3 phishscore=0 adultscore=0 bulkscore= classifier=spam adjust=0 reason=mlx scancount=1 engine=7.0.1-1308280000 definitions=main-131008006

HACKTIVITY

Old fashioned sources
„client IP”

Tracing route to mail-we0-f182.google.com [74.125.82.182] over a maximum of 30 hops:

1	17 ms	2 ms	*	192.168.4.1
2	17 ms	17 ms	17 ms	192.168.0.1
3				
4	11 ms	26 ms	17 ms	catv-89-135-222-30.catv.broadband.hu [89.135.222.30]
5	17 ms	17 ms	17 ms	84.116.240.78
6	17 ms	17 ms	40 ms	84.116.240.1 - **UPC Austria GmbH**
7	**17 ms**	**17 ms**	**14 ms**	72.14.217.146 - **Google Inc**
8	27 ms	17 ms	41 ms	209.85.243.121
9	47 ms	48 ms	48 ms	72.14.234.11
10	40 ms	55 ms	63 ms	209.85.241.228
11	52 ms	62 ms	56 ms	209.85.240.221
12	62 ms	58 ms	39 ms	209.85.252.83
13	*	*	*	Request timed out.
14	50 ms	58 ms	43 ms	mail-we0-f182.google.com [74.125.82.182]

Trace complete.

Old fashined sources - email
….hipotetical interpretation

- the Google does'nt make transatlantic traffic if not neccessary!
- tipical ping times between Europe and USA :)
- „de novo” signal
- syllogism: the server and the sender both located near Vienna +/- 200 km
- (dedicated connection across Google and ISP?)

HACKTIVITY

Old fashioned sources
DNS-records

- non-masked A-record – note: entire IP-block may owned by one server/owner as „hosting provider”
- masked A-record – eg. Cloudflare:
  - by default adding direct-connect subdom.
  - cannot masking MX – reverse lookup!

HACKTIVITY

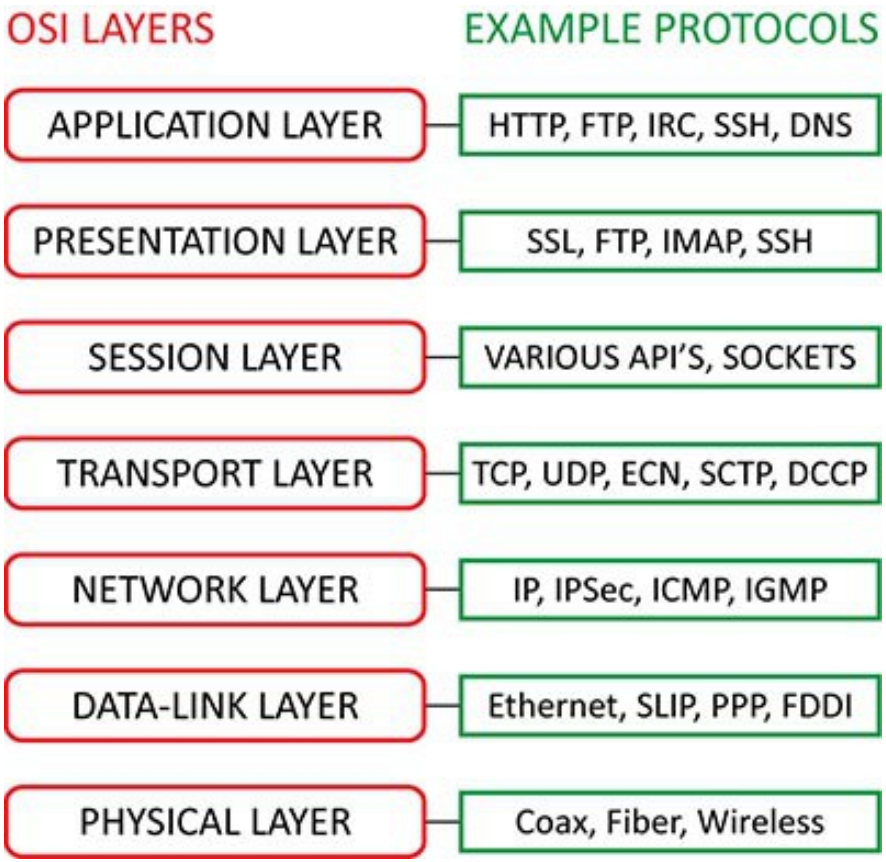
Old fashioned sources
cookies and HTTP-responses

- same owner - same developer - same admin - same CMS - frequent
- consequence – similar HTTP-stream patterns and cookie-patterns between suspected websites

HACKTIVITY

Different layers, different methods

…and cognitive/linguistical/behavioural



Question:
how can examine...

...and why?

fig: http://essayweb.net/miscellany/datatransmission.shtml

HACKTIVITY

Why we loves the Facebook API?
Facebook Graph API (2013. sept.)

- Graph API object set: Achievement, Album, Application, Checkin, **Comment**, Domain, Errors, Event, FriendList, Group, Insights, Link, Message, **Note**, Offer, Order, Page, Payment, Photo, Pictures, **Post**, Question, QuestionOption, Review, **Status message**, Thread, User, Video

- The fields of the User object:

id, name, first_name, middle_name, last_name, **gender**, locale, languages, link, username, **age_range**, third_party_id, installed, timezone, updated_time, verified, **bio**, birthday, cover, currency, devices, education, email, hometown, interested_in, location, political, payment_pricepoints, favorite_athletes, favoriate_teams, picture, quotes, relationship_status, religion, security_settings, segnificant_other, video_upload_limits, website, work

HACKTIVITY

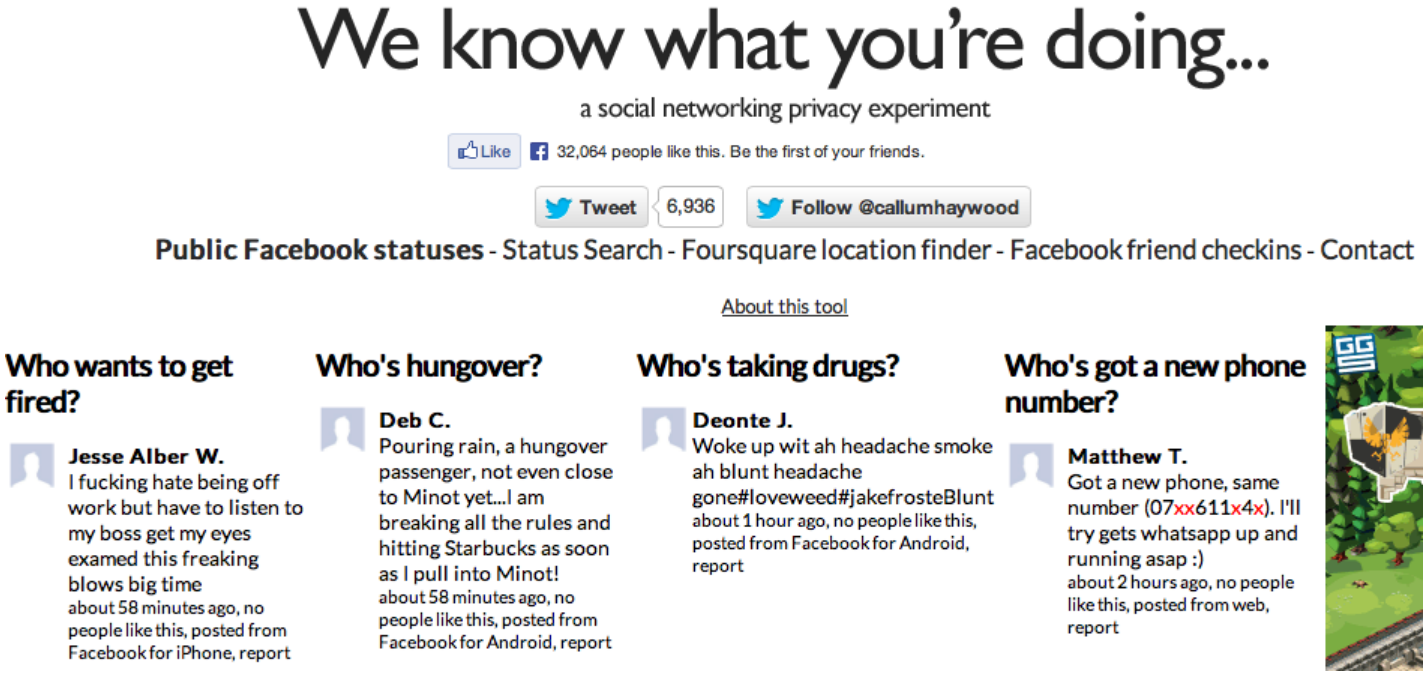
Why we loves social sources?

Some methods not needs special permissions, API keys, etc. Eg. this simple URL:

graph.facebook.com/https://facebook.com/hacktivity?metadata=1

- In real e.g.

http://www.weknowwhatyouredoing.com/



HACKTIVITY

The rabbit hole
one of my fav. studies

Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)

- input: 58000 volunteers likes, some demographical and psychometrical date
- Processed with simple linear regression and logistics regression

HACKTIVITY

The found, successfully like-correlated properties:

"sexual orientation," "ethnic origin," "political views," "religion," "personality," "intelligence," "satisfaction with life" (SWL), substance use ("alcohol," "drugs," "cigarettes"), "whether an individual's parents stayed together until the individual was 21 y old," and basic demographic attributes such as "age," "gender," "relationship status," and "size and density of the friendship network."

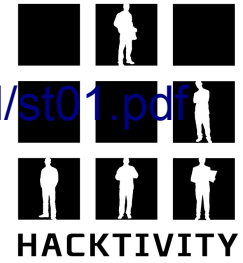
Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)

The researcher...

Trait		Selected most predictive Likes		
IQ	*High*	The Godfather	Jason Aldean	*Low*
		Mozart	Tyler Perry	
		Thunderstorms	Sephora	
		The Colbert Report	Chiq	
		Morgan Freemans Voice	Bret Michaels	
		The Daily Show	Clark Griswold	
		Lord Of The Rings	Bebe	
		To Kill A Mockingbird	I Love Being A Mom	
		Science	Harley Davidson	
		Curly Fries	Lady Antebellum	
Satisfaction With Life	*Satisfied*	Sarah Palin	Hawthorne Heights	*Dissatisfied*
		Glenn Beck	Kickass	
		Proud To Be Christian	Atreyu (Metal Band)	
		Indiana Jones	Lamb Of God	
		Swimming	Gorillaz	
		Jesus Christ	Science	
		Bible	Quote Portal	
		Jesus	Stewie Griffin	
		Being Conservative	Killswitch Engage	
		Pride And Prejudice	Ipod	

http://www.pnas.org/content/suppl/2013/03/07/1218772110.DCSupplemental/st01.pdf

Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)



...can...

Openness	*Liberal & Artistic*	Oscar Wilde Charles Bukowski Sylvia Plath Leonardo Da Vinci Bauhaus Dmt The Spirit Molecule American Gods John Waters Plato Leonard Cohen	NASCAR Austin Collie Monster-In-Law I don't read Justin Moore ESPN2 Farmlandia The Bachelor Oklahoma State University Teen Mom 2	*Conservative*
Conscientiousness	*Well Organized*	Law Officer National Law Enforcement Lowfares.Com Accounting Foursquare Emergency Medical Services Sunday Best Kaplan University Glock Inc Mycalendar 2010	Wes Anderson Bandit Nation Omegle Vocaloid Serial Killer Screamo Anime Vamplets Join If Ur Fat Not Dying	*Spontaneous*
Extraversion	*Outgoing & Active*	Beerpong Michael Jordan Dancing Socializing Chris Tucker I Feel Better Tan Modeling Cheerleading Theatre Flip Cup	RPGs Fanfiction.Net Programming Anime Manga Video Games Role Playing Games Minecraft Voltaire Terry Pratchet	*Shy & Reserved*

http://www.pnas.org/content/suppl/2013/03/07/1218772110.DCSupplemental/st01.pdf

Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)

HACKTIVITY

...read between...

Agreeableness	*Cooperative*	Compassion International	I Hate Everyone	*Competitive*
		Logan Utah	I Hate You	
		Jon Foreman	I Hate Police	
		Redeeming Love	Friedrich Nietzsche	
		Pornography Harms	Timmy South Park	
		The Book Of Mormon	Atheism / Satanism	
		Circles Of Prayer	Prada	
		Go To Church	Sun Tzu	
		Christianity	Julius Caesar	
		Marianne Williamson	Knives	
Emotional Stability	*Neurotic*	Sometimes I Hate Myself	Business Administration	*Calm & Relaxed*
		Emo	Getting Money	
		Girl Interrupted	Parkour	
		So So Happy	Track & Field	
		The Addams Family	Skydiving	
		Vocaloid	Mountain Biking	
		Sixbillionsecrets.com	Soccer	
		Vampires Everywhere	Climbing	
		Kurt Donald Cobain	Physics / Engineering	
		Dot Dot Curve	48 Laws Of Power	
Gender	*Female*	Tv Fanatic	Modern Warfare 2	*Male*
		Chiq	ESPN	
		Gillette Venus	Sportscenter	
		Shoedazzle	Band Of Brothers	
		Bebe	Starcraft	
		Proud To Be A Mom	Deadliest Warrior	
		Covergirl	Dos Equis	
		Wet Seal	Red Vs Blue	
		Aerie By American Eagle	X Games	
		Mall World	Bruce Lee	

http://www.pnas.org/content/suppl/2013/03/07/1218772110.DCSupplemental/st01.pdf

Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)

HACKTIVITY

...the ~~lines~~ likes

Politics	*Republican*	George W Bush John McCain Conservative Rush Limbaugh Sean Hannity Bill Oreilly Positively Republican Sarah Palin Ronald Reagan Glenn Beck	Joe Biden Speaker Nancy Pelosi Health Care Reform The White House Democrats Barbara Boxer Anthony Weiner Being Liberal Left Action Barack Obama2012 Ted Kennedy	*Democrat*
Sexual Orientation	*Homosexual Males*	No H8 Campaign Kathy Griffin Kurt Hummel Glee Human Rights Campaign Mac Cosmetics Adam Lambert Ellen DeGeneres Juicy Couture Sue Sylvester Glee Wicked The Musical	X Games Nike Basketball Bungie WWE Sportsnation Wu-Tang Clan Foot Locker Shaq Bruce Lee Being Confused After Waking Up From Naps	*Heterosexual Males*
	Homosexual Females	Girls Who Like Boys Who Like Boys Rupauls Drag Race No H8 Campaign Gay Marriage Human Rights Campaign The L Word Sometimes I Just Lay In Bed And Think About Life Not Being Pregnant Gay Marriage Tegan And Sara	Lipton Brisk Yahoo Adidas Originals Foot Locker WWE Inbox 1 Makes Me Nervous Thinking Of Something And Laughing Alone I Just Realized Immature Spells I'm Mature Did You Get A Haircut No It Grew Shorter	*Heterosexual Females*

http://www.pnas.org/content/suppl/2013/03/07/1218772110.DCSupplemental/st01.pdf

Private traits and attributes are predictable from digital records of human behavior (Michal Kosinskia, David Stillwella, Thore Graepelb, 2012)

HACKTIVITY

Twitter-users closer

- **problem:** e.g. linkfarm identification
- **hipothesis:** Let's measure the influence and popularity.

Twitter
Similis simili gaudet

- popular users have more follower than followed
- deliberate connections
- identification of problematic accounts [bots/linkfarms]: measure the popularity and the followers dispersion. Abnormal dispersion may indicate problematic user.
- (cluster analysis)

Twitter

Measure the most popular followers of the suspected user

- An elegant implementation in Python
- sample result:

droidPisti 12
FakeFeri 11
HirhozoHugo 9
Blogcsokor 9
Wannabehirportal 8….

Twitter
Entity analysis in the stream

- *Extracting entities from tweets and performing simple frequency analysis - (Mining the Social Web)*
- Abnormal dispersion simple result:

#kiralywebhely 2834
#legujabbpinamagazin 2830
#ittmindenwareztmegtalasz 2820
#torrentekmindenmennyisegben 2811

HACKTIVITY

Possibilities of IF-IDF measures

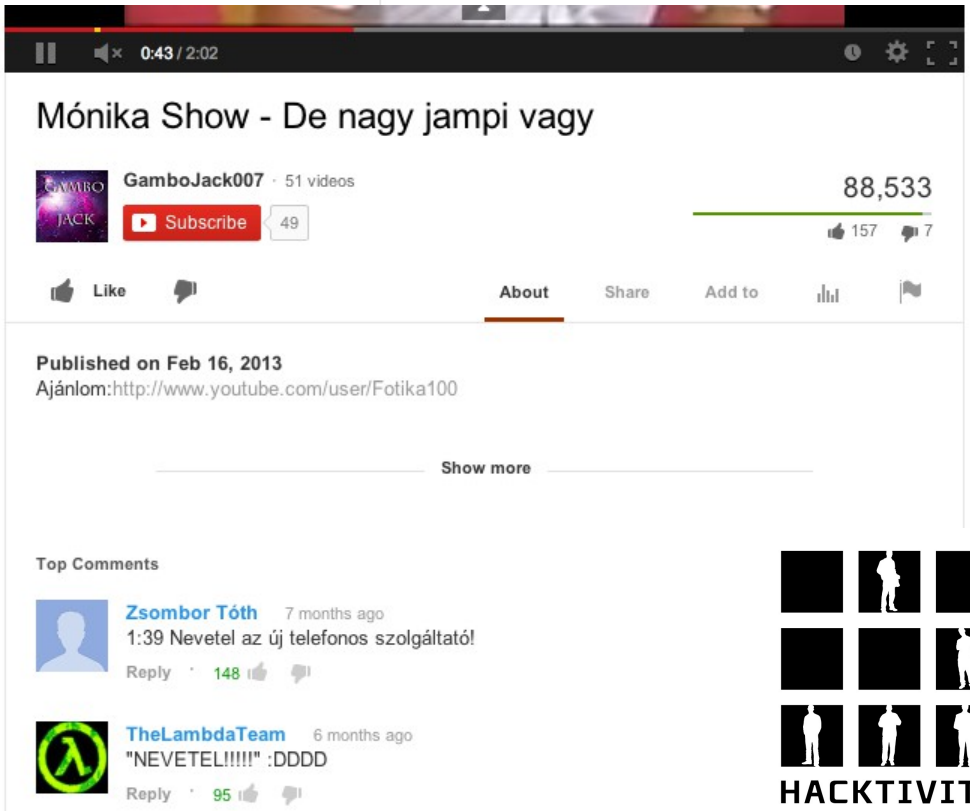
About digital humanities and computational linguistics

Example: linguistics and the fine tuned search

„To be, or not to be.”

Shakespeare W. (soap opera addict), cca. 1603

An identified former YT-hapax legomenon and the „did you mean”: *nevetel*-novotel

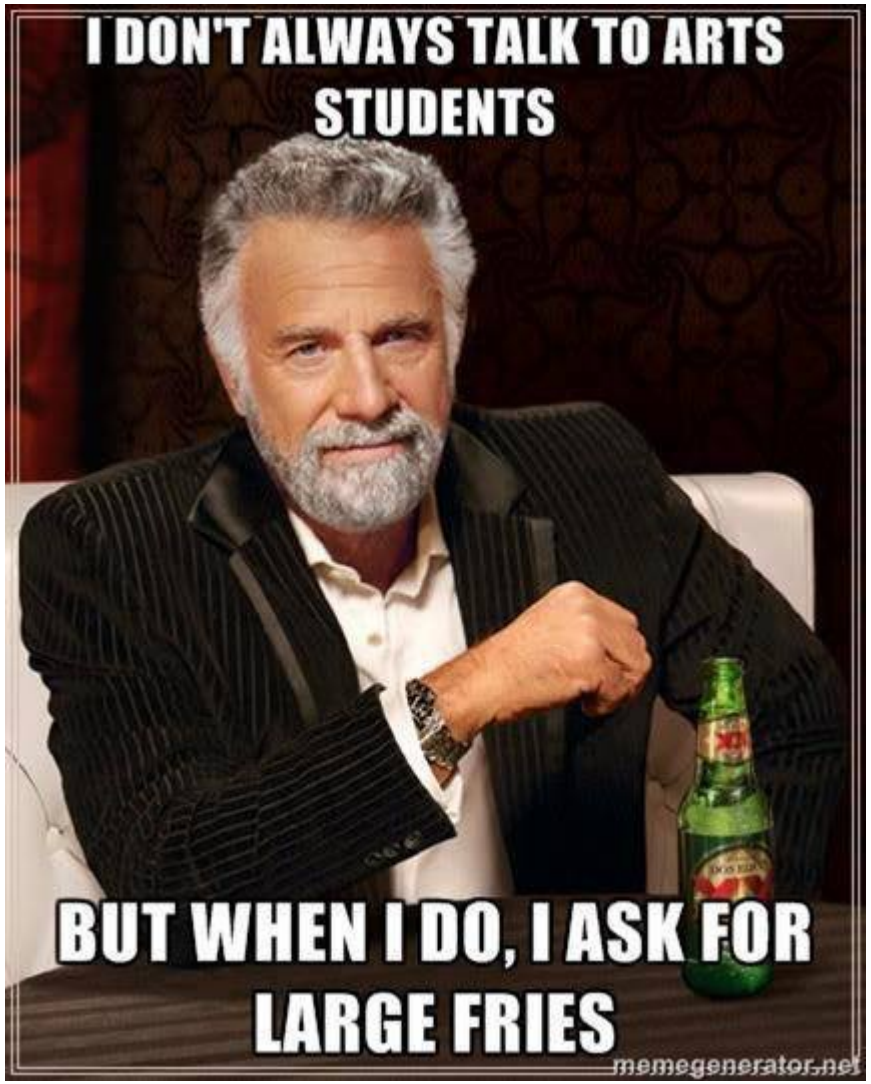


In this case, the titles of the first two results, don't contains „nevetel”, only the comments (with significant weight)

„Jampi” Zsolti, 2009,
Maunika Erdélyi et al.

HACKTIVITY

Who they are really? - popular misbelief



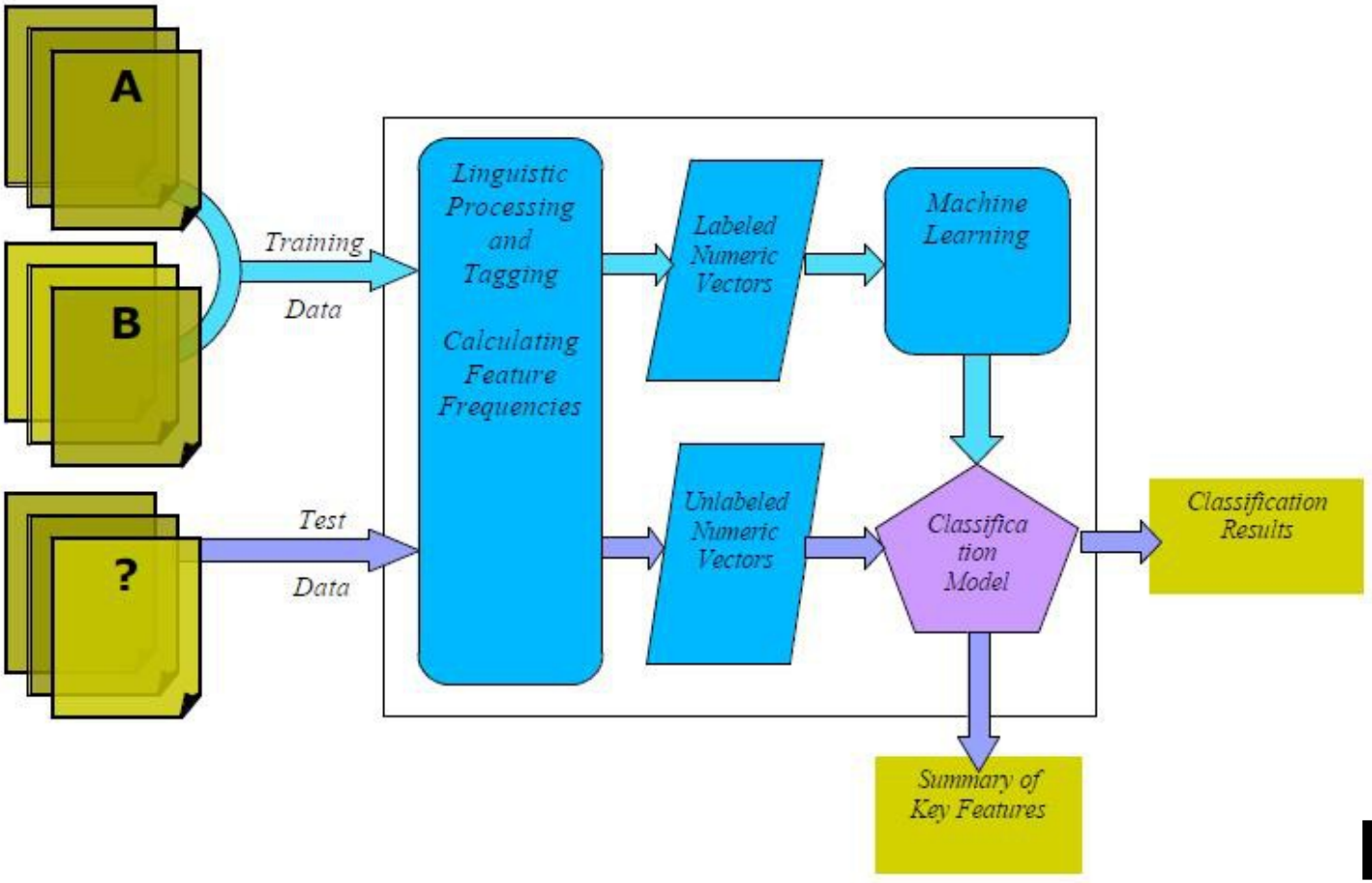
HACKTIVITY

Computational linguists: they are not simply BAs/MAs. They must have experiences in...

strong mathematical and computer science skills, such as

- classifiers
- machine learning
- declarative programming
- fields of AI
- data mining

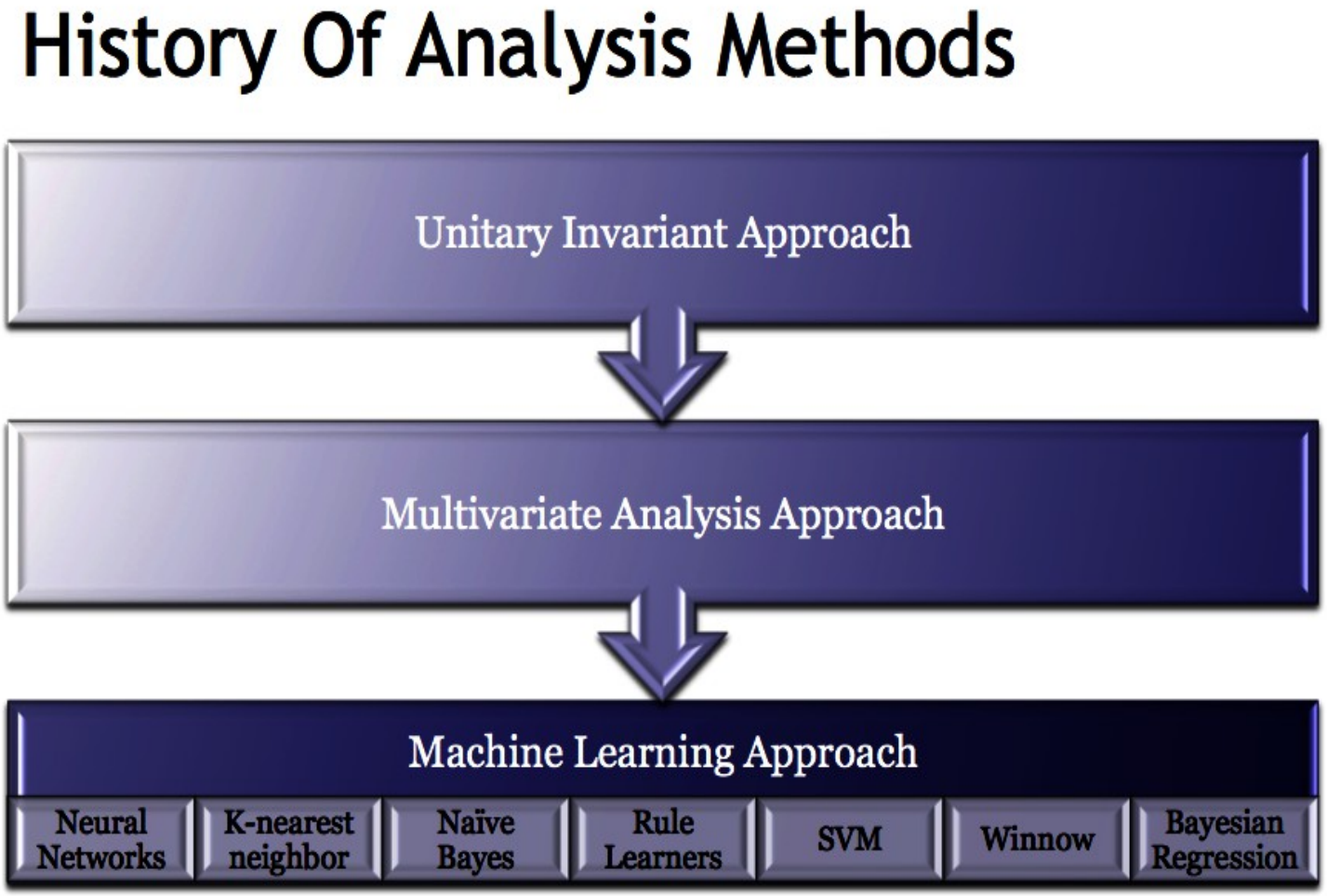
One principle of the many



Automatically Profiling the Author of an Anonymous Text
(Shlomo Argamon, Moshe Koppel, James W. Pennebaker, Jonathan Schler)

HACKTIVITY

An quick historical overview

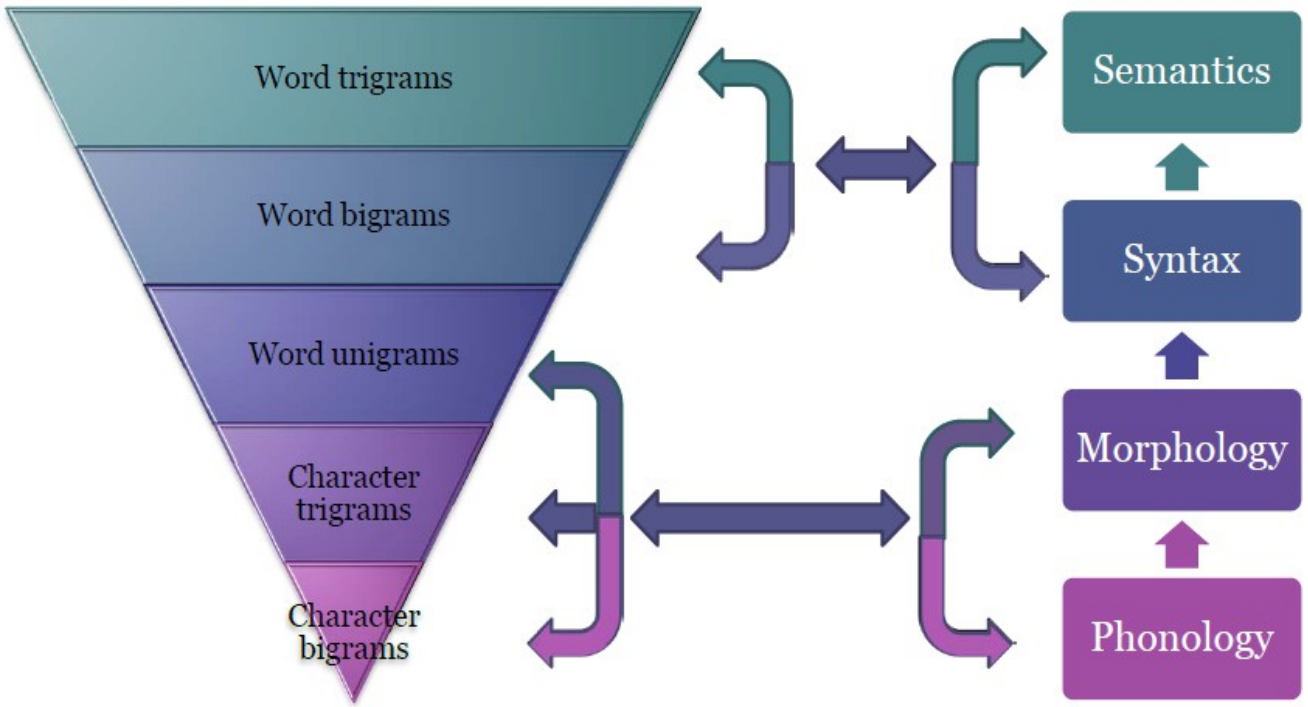


Computational Methods in Authorship Attribution,
Daniel Labenski, NLP Seminar 2012

HACKTIVITY

Authorship identification on big sets

An hierarchical representation of features and related linguistic levels



Authorship identification in large email collections: Experiments using features that belong to different linguistic levels (George K. Mikros & Kostas Perifanos, 2011)

HACKTIVITY

About Natural Language Processing Python NLTK

Typical:

1. End of Sentence (EOS) Detection
2. Tokenization
3. Part-of-Speech Tagging
4. Chunking
5. Extraction

```
nltk.tokenize.sent_tokenize(valtozo)
tokenek = [nltk.tokenize.word_tokenize(s) for s in sentences]
 pos_tagged_tokens = [nltk.pos_tag(t) for t in tokenek]
 pos_tagged_tokens
chunking
mondattani elemzés -
 ne_chunks = nltk.batch_ne_chunk(pos_tagged_tokens)
 ne_chunks
```

Some selected linguistics weapons in authorship identification and plagriarism detection

- frequency and dispersion of suffixed words
- Syntactic complexity
  - Syntactic tree depth
  - Distance between syntactically dependent words
- cohesion

A multitude of linguistically-rich features for authorshipattribution (Ludovic Tanguy, Assaf Urieli, Basilio Calderone, Nabil Hathout, Franck Sajous, 2011)

HACKTIVITY

The style of the sentences likes the authors fingerprint!

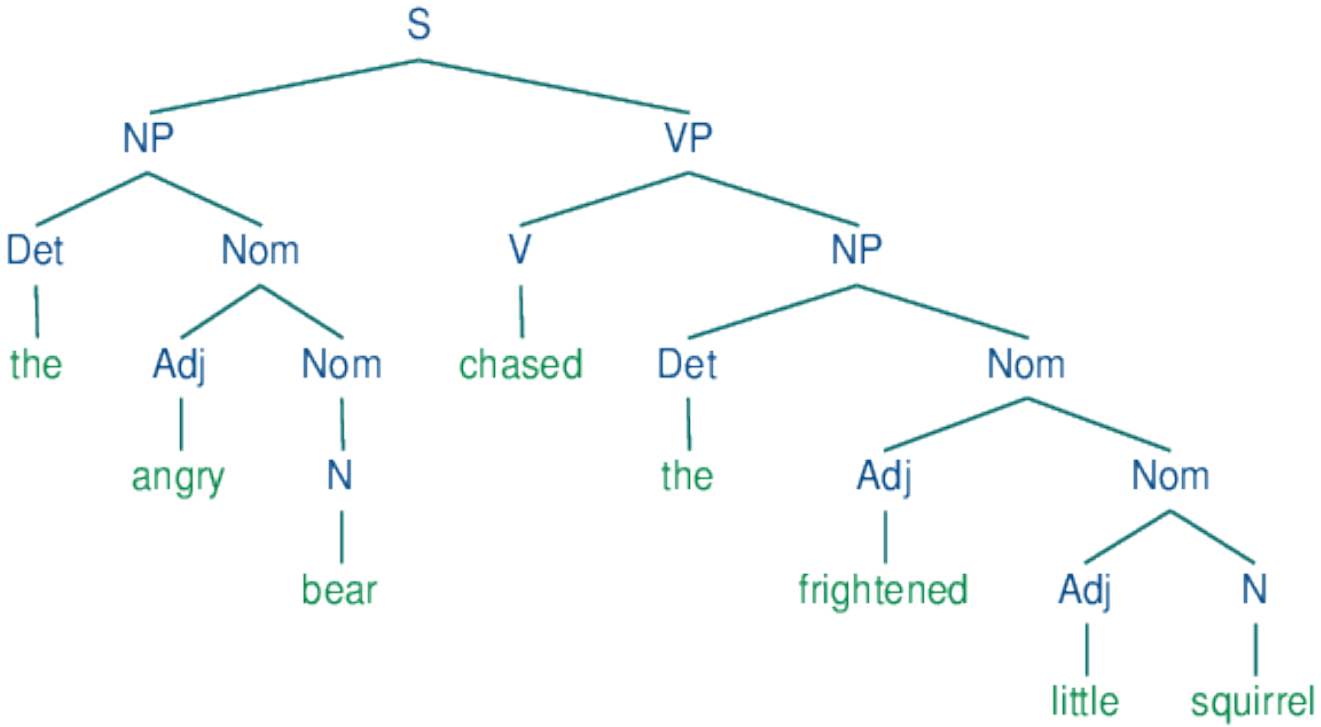


Fig: http://nltk.org/book/ch08.html

HACKTIVITY

Quick explanation of a plagiarism detection technique [translation plagiarism]

core problems:

- non-friendly language pairs:
a) fixed/non-fixed word order, b) conjugation, c) maior grammar differences

- former, partially solutions

Principles of new algo:
- sentence as translation unit
- sentence as context
- sentence, because
1. represent an unit in mind
2. easy to find the borders
3. enought unique
-similarity metrics

Many-many sources, mostly from Máté Pataki (MTA SZTAKI)
http://www.inf.u-szeged.hu/mszny2011/

HACKTIVITY

Quick explanation of a plagiarism detection technique [translation plagiarism] – algo in nutshell (similarity metrics)

Let's $w_x \in S_x$ and $w_y \in S_y$ $\quad Sim(x,y) = \alpha \cdot |S_x \cap S_y| - \beta \cdot |S_x \setminus S_y|$
will show the found and not-found words (with different weights).

After a bit fine tunin $Sim(x,y) = \min(\alpha \cdot |S_x \cap S_y| - \beta \cdot |S_x \setminus S_y|, \alpha \cdot |S_y \cap S_x| - \beta \cdot |S_y \setminus S_x|)$

Some lemmas $trans(w_x) = W_y$ where $w_y \in W_y$

$trans(w_y) = W_x$ where $w_x \in W_x$

$w_x \in trans(w_y)$ therefore $w_y \in trans(w_x)$

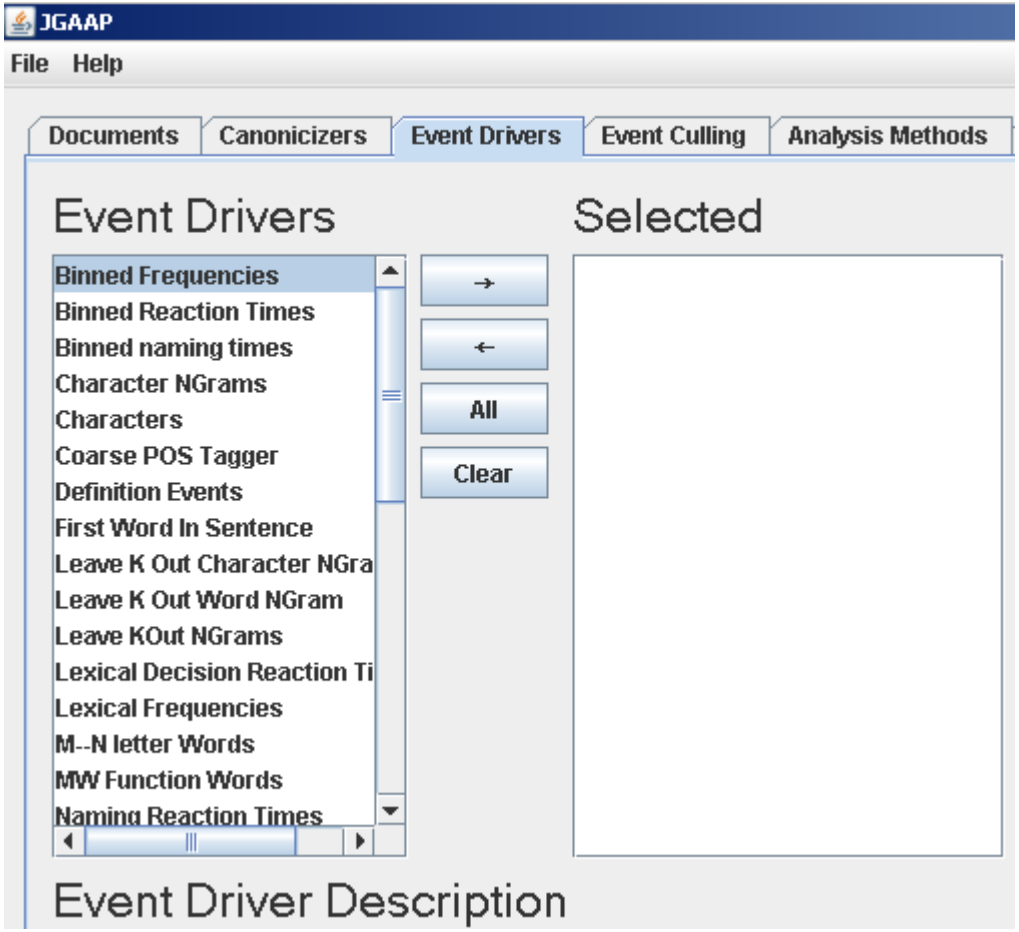
$w_y \in trans(w_x)$ therefore $w_x \equiv w_y$

Signs: S – sentence, n, w – number of words in sentence and words

Many-many sources, mostly from Máté Pataki (MTA SZTAKI)

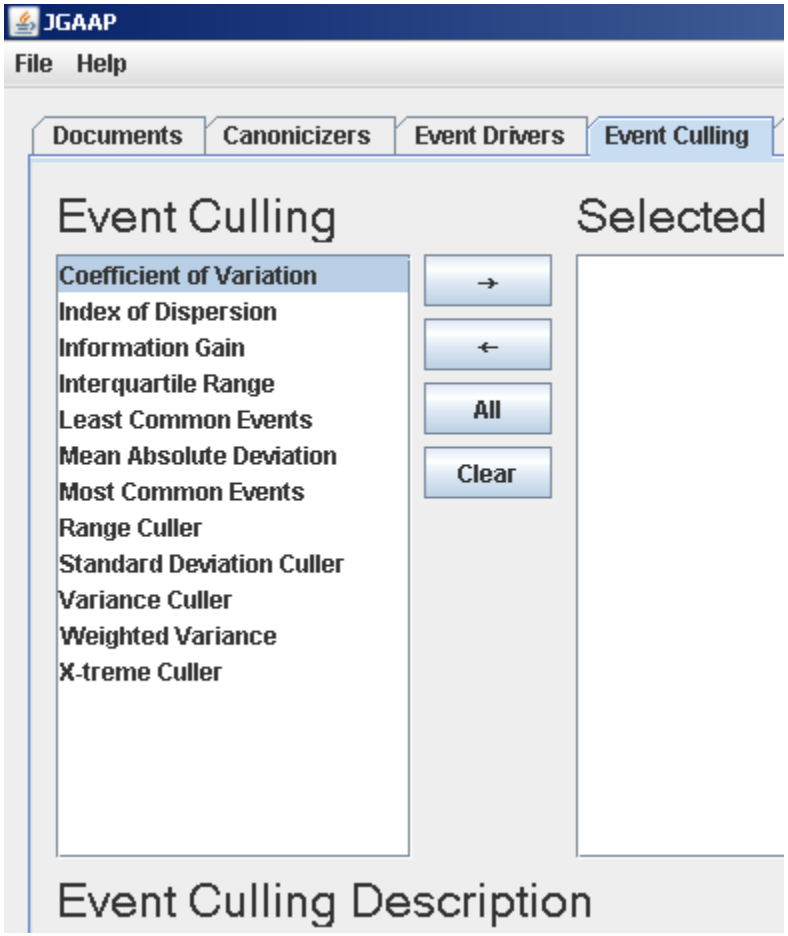
HACKTIVITY

Go deep – JGAAP, event drivers



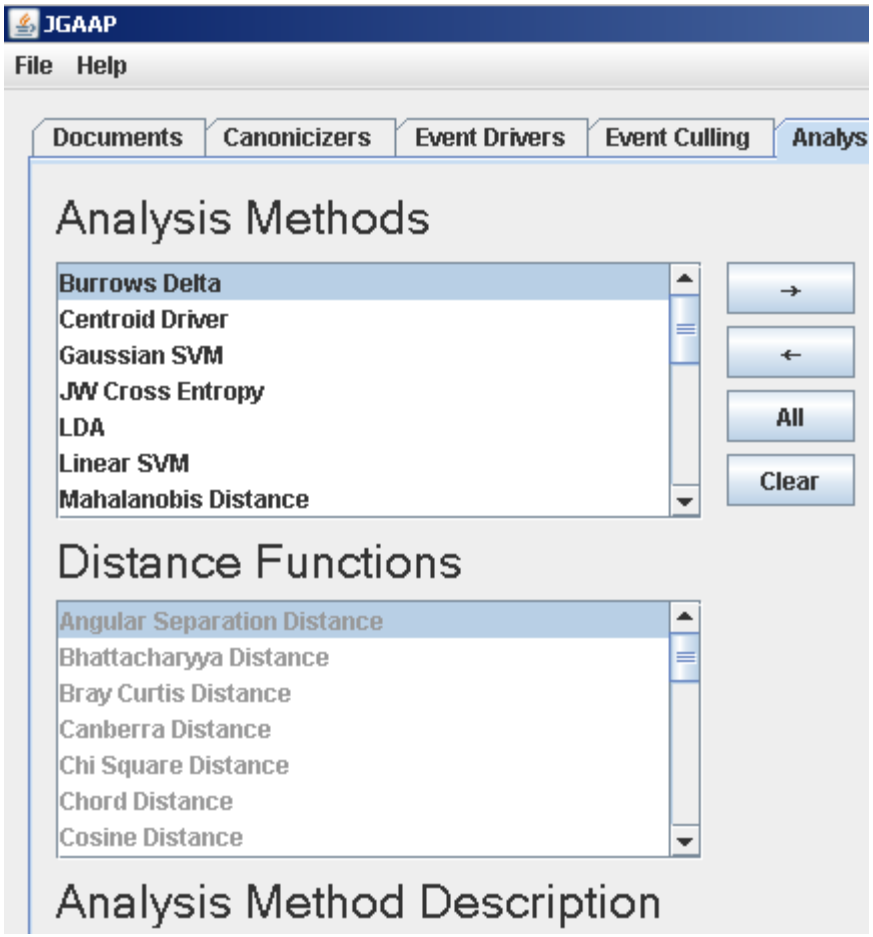
HACKTIVITY

Go deep – JGAAP, event culling



HACKTIVITY

Go deep – JGAAP, analysis methods



HACKTIVITY

See also

- Text Mining: Applications and Theory (Michael W. Berry, Jacob Kogan)
- Natural Language Processing with Python (Steven Bird, Ewan Klein, Edward Loper)
- Mining the Social Web (Matthew A. Russel)
- Forensic Linguistics - Advances in Forensic Stylistics (Gerald McMenamin et al.)
- The Routledge Handbook of Forensic Linguistics (ed. Malcolm Coulthard, Alison Johnson)

PAN CLEF papers

Recommended blogs in hungarian:
http://kereses.blog.hu/
http://kep-es-kod.blogspot.ch/
http://szamitogepesnyelveszet.blogspot.ch/

- **Digital humanities MA in Hungary:**
  http://www.btk.unideb.hu/
  https://btk.ppke.hu/

HACKTIVITY