# Tko je to napisao?

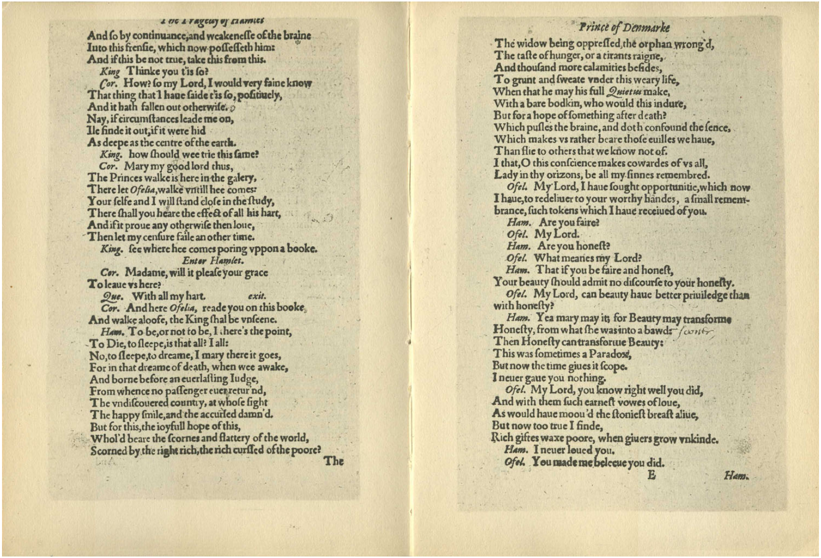
## Analiza autorstva metodama računalne lingvistike

Jan Šnajder

TakeLab FER
Sveučilište u Zagrebu

Centar informacijske sigurnosti FER-a
23. studenog 2016.

# Tekst, tekst, tekst



*The Tragedy of Hamlet*

And ſo by continuance,and weakeneſſe of the braine
Into this frenſie, which now poſſeſſeth him:
And if this be not true, take this from this.
*King* Thinke you t'is ſo?
*Cor.* How? ſo my Lord, I would very faine know
That thing that I haue ſaide t'is ſo, poſitiuely,
And it hath fallen out otherwiſe.
Nay, if circumſtances leade me on,
Ile finde it out,if it were hid
As deepe as the centre of the earth.
*King.* how ſhould wee trie this ſame?
*Cor.* Mary my good lord thus,
The Princes walke is here in the galery,
There let *Ofelia*,walke vntill hee comes:
Your ſelfe and I will ſtand cloſe in the ſtudy,
There ſhall you heare the effect of all his hart,
And if it proue any otherwiſe then loue,
Then let my cenſure faile an other time.
*King.* ſee where hee comes poring vppon a booke.
*Enter Hamlet.*
*Cor.* Madame, will it pleaſe your grace
To leaue vs here?
*Que.* With all my hart. *exit.*
*Cor.* And here *Ofelia*, reade you on this booke,
And walke aloofe, the King ſhal be vnſeene.
*Ham.* To be,or not to be, I there's the point,
To Die, to ſleepe,is that all? I all:
No,to ſleepe,to dreame, I mary there it goes,
For in that dreame of death, when wee awake,
And borne before an euerlaſting Iudge,
From whence no paſſenger euer retur'nd,
The vndiſcouered country, at whoſe ſight
The happy ſmile,and the accurſed damn'd.
But for this,the ioyfull hope of this,
Whol'd beare the ſcornes and flattery of the world,
Scorned by the right rich,the rich curſſed of the poore?
The

*Prince of Denmarke*

The widow being oppreſſed,the orphan wrong'd,
The taſte of hunger, or a tirants raigne,
And thouſand more calamities beſides,
To grunt and ſweate vnder this weary life,
When that he may his full *Quietus* make,
With a bare bodkin, who would this indure,
But for a hope of ſomething after death?
Which puſles the braine, and doth confound the ſence,
Which makes vs rather beare thoſe euilles we haue,
Than flie to others that we know not of.
I that,O this conſcience makes cowardes of vs all,
Lady in thy orizons, be all my ſinnes remembred.
*Ofel.* My Lord, I haue ſought opportunitie,which now
I haue,to redeliuer to your worthy handes, a ſmall remem-
brance, ſuch tokens which I haue receiued of you.
*Ham.* Are you faire?
*Ofel.* My Lord.
*Ham.* Are you honeſt?
*Ofel.* What meanes my Lord?
*Ham.* That if you be faire and honeſt,
Your beauty ſhould admit no diſcourſe to your honeſty.
*Ofel.* My Lord, can beauty haue better priuiledge than
with honeſty?
*Ham.* Yea mary may it; for Beauty may transforme
Honeſty, from what ſhe was into a bawd:
Then Honeſty can transforme Beauty:
This was ſometimes a Paradox,
But now the time giues it ſcope.
I neuer gaue you nothing.
*Ofel.* My Lord, you know right well you did,
And with them ſuch earneſt vowes of loue,
As would haue moou'd the ſtonieſt breaſt aliue,
But now too true I finde,
Rich giftes waxe poore, when giuers grow vnkinde.
*Ham.* I neuer loued you.
*Ofel.* You made me beleeue you did.
E *Ham.*

# Tekst, tekst, tekst

---

**From:** H <hrod17@clintonemail.com>
**Sent:** Wednesday, September 12, 2012 9:12 PM
**To:** Diane Reynolds
**Subject:** Re:

I'm home and up for another hour if you can talk now. B6

----- Original Message -----
From: Diane Reynolds
Sent: Wednesday, September 12, 2012 06:26 PM
To: H
Subject:

I am so sorry and sad about all of what has and is happening in Cairo, Benghazi and elsewhere in the ME and beyond. Just called your office to tell you and heard you're at the WH so emailing.

B6

# Forenzička lingvistika

- **Atribucija autorstva**: Tko je autor?
- **Provjera autorstva**: Je li X autor?
- **Profiliranje autora**: Kakav je autor?
- **Otkrivanje plagijata**: Je li tekst prepisan?

# Primijenjena lingvistika

- Rješenja praktičnih probleme povezanih s jezikom
- Interdisciplinarna
- **Forenzička lingvistika**
  - lingvističke metode u kontekstu forenzike (pravo, jezik, kriminalistika)
- **Stilistika**
  - proučavanje jezičnog odnosno književnog stila

# Forenzička lingvistika

# Stilistika

- **Jezična varijacija** je temeljna karakteristika jezika
  - fonologija, leksikon, gramatika
- Ključni koncept **sociolingvistike**
  - lingvistička varijacija $\Leftrightarrow$ društvene karakteristike
- **Forenzička stilistika**
  - stil karakterističan za pojedinca (idiolekt)
- **Stilometrija**
  - statističke i računalne metode primjene stilistike

# Forenzička lingvistika – primjene

- **Kibernetički kriminal**
  - phishing scams, spam, ucjene, uznemiravanje
  - SMS, e-pošta, blogovi
- **Marketing i društvena istraživanja**
  - karakteristike korisnika društvenih mreža
  - demografske značajke, političke/potrošačke preferencije
- **Znanost o književnosti i obrazovanje**
  - utvrđivanje kontroverznog autorstva, kvalitete prijevoda, osobine ličnosti studenata
  - detekcija plagijata u akademskim publikacijama

# Plan

1. NLP i strojno učenje
2. Atribucija autorstva
3. Provjera autorstva
4. Profiliranje autora

# Plan

1. NLP i strojno učenje
2. Atribucija autorstva
3. Provjera autorstva
4. Profiliranje autora

# Računala i lingvistika

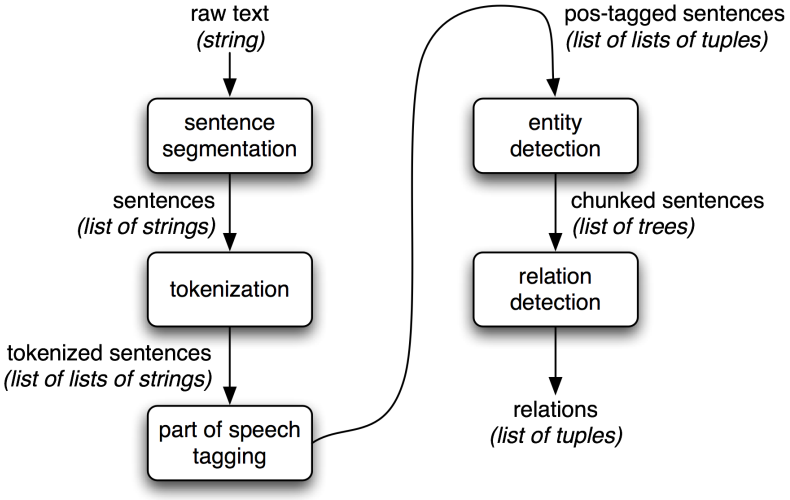
- **Računalna lingvistika**
  - “znanstveno istraživanje jezika iz računalne perspektive. . . zainteresirana za računalne modele jezičnih fenomena” (ACL)
- **Obrada prirodnog jezika (NLP)**
  - područje računarske znanosti i umjetne inteligencije koje se bavi interakcijom čovjeka i računala kroz prirodne (ljudske) jezike

$\Rightarrow$ Računalna forenzička lingvistika

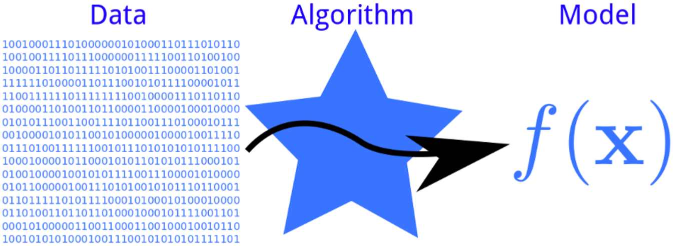
# Tipični zadatci

- Morfološka analiza/segmentacija
- Označavanje vrste riječi
- Parsanje (sintaktička analiza)
- Razrješavanje višeznačnosti riječi
- Razrješavanje koreferencije
- Prepoznavanje imenovanih entiteta
- Strojno prevođenje
- ...

# Tipični koraci



# Strojno učenje




- Algoritmi za (polu)automatsku ekstrakciju novog i korisnog znanja – u obliku pravila, uzoraka ili modela – iz proizvoljnih skupova podataka

# Strojno učenje i NLP

- Za zadani ulaz, algoritam (**klasifikator**) dodijeljuje odluku (najčešće **da/ne**)
- Velik broj problema u NLP-u može se svesti na donošenje odluke ili niz odluka
- Verifikacija autorstva: za zadani ulazni tekst, odluči je li X autor (da/ne)
- Atribucija autorstva: za zadani ulazni tekst, odluči tko je autor (odluka iz skupa opcija)

# Primjena modela strojnog učenja

1. Priprema podataka
2. Ekstrakcija značajki
3. Učenje (treniranje) modela
4. Evaluacija
5. Dijagnosticiranje
6. Ugradnja

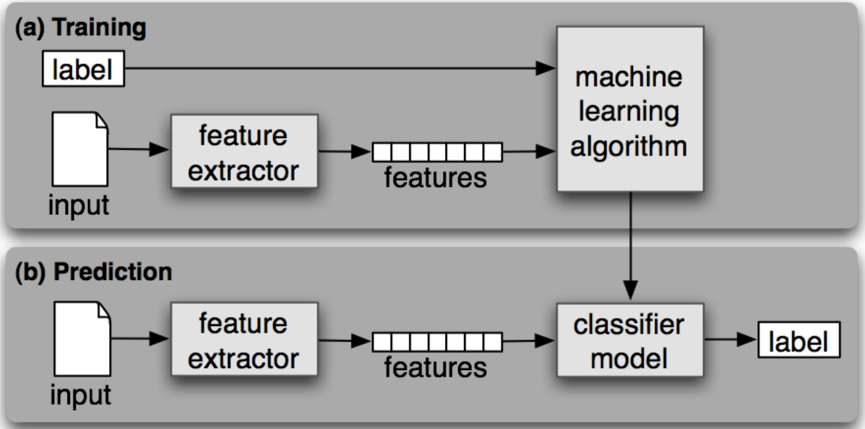
# Pristupi

- **Nadzirano (supervised)**
  - klasifikacija
  - regresija
  - učenje rangiranja (*learning to rank*)
- **Nenadzirano (unsupervised)**
  - grupiranje (*clustering*)
  - novelty/outlier detection

# Predikcija

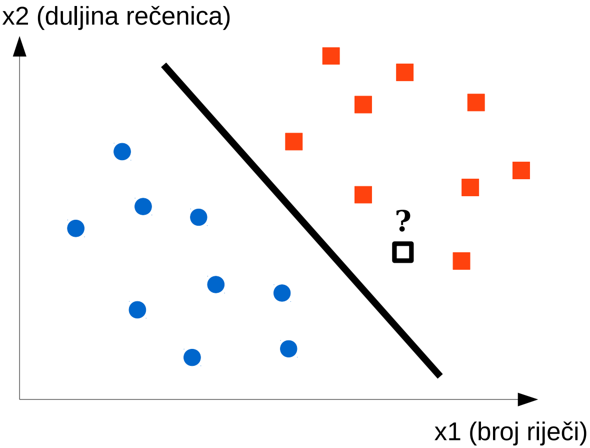
- Model na temelju viđenih podataka zaključuje nešto o novim podatcima
- Model mora moći **generalizirati**
- Naš cilj: napraviti model koji dobro generalizira

# Nadzirano učenje

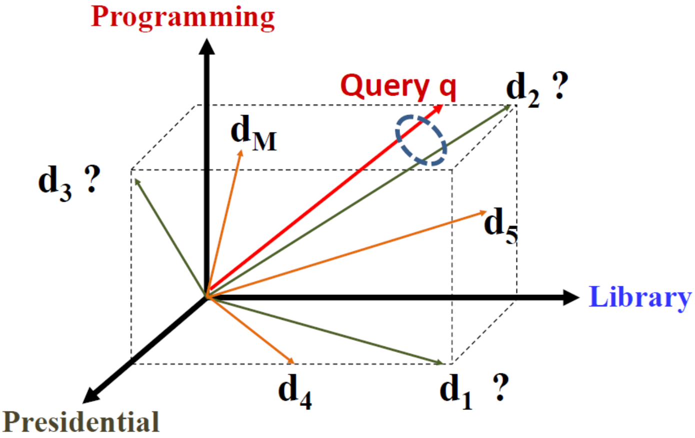



http://www.nltk.org/book/ch06.html

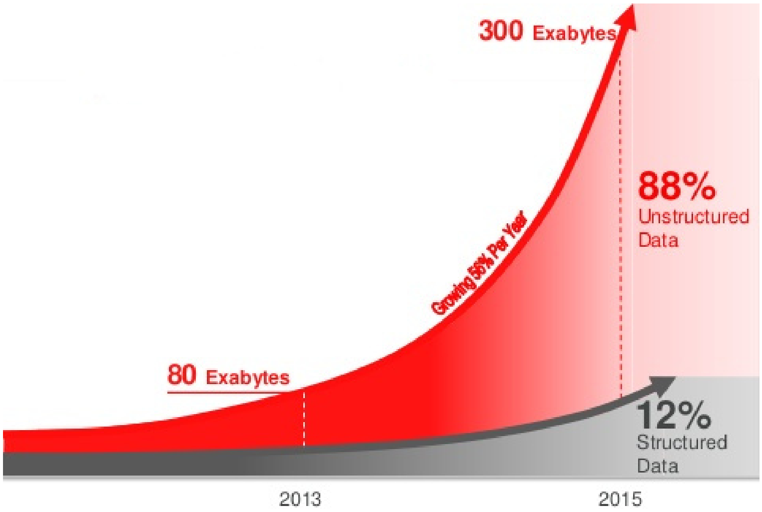
# Klasifikacija



# Vektorski model dokumenta



# Zašto sada?



300 Exabytes
Growing 56% Per Year
88%
Unstructured Data
80 Exabytes
12%
Structured Data
2013
2015

# Znanost o podatcima

# Plan

1. NLP i strojno učenje
2. Atribucija autorstva
3. Provjera autorstva
4. Profiliranje autora

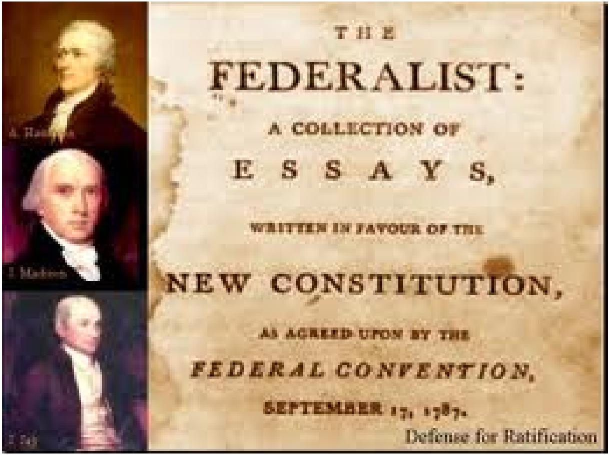
# Izvori


- Stamatatos, Efstathios. **A survey of modern authorship attribution methods**. Journal of the American Society for information Science and Technology 60.3 (2009): 538-556.
- Koppel, M., Schler, J., & Argamon, S. (2009). **Computational methods in authorship attribution**. Journal of the American Society for information Science and Technology, 60(1), 9-26.

# Povijest

- Srednji vijek: autorstvo = istinitost teksta
- Jedna invarijantna značajka
  - Mendenhall (1887): Shakespeare, Bacon, Marlowe
- Multivarijatna analiza
  - Mosteller and Wallace (1964): "The Federalist Papers"
  - Naivni Bayes i više značajki

# The Federalist Papers




85 eseja zagovornika američkog ustava iz 1787.:
Alexander Hamilton, James Madison, John Jay

# Povijest

- **1964–1990**
    - definiranje stilometrijskih značajki
    - više od 1000 različitih mjera do kraja 1990.
    - problem: evaluacija
- **1990–danas**
    - strojno učenje i NLP-a (klasifikacija teksta)
    - velike količine tekstova na internetu
    - obavještajstvo, kriminalistika, pravo
    - objektivna i standardizirana evalucija

# Atribucija autorstva strojnim učenjem

- Problem **višeklasne klasifikacije teksta**
- Iskorištavanje velikog broja potencijalnog korisnih tekstnih (stilometrijskih) **značajki**
- Postupci odabira značajki

# Stilometrijske značajke

| Features | | Required tools and resources |
|---|---|---|
| Lexical | Token-based (word length, sentence length, etc.) | Tokenizer, [Sentence splitter] |
| | Vocabulary richness | Tokenizer |
| | Word frequencies | Tokenizer, [Stemmer, Lemmatizer] |
| | Word *n*-grams | Tokenizer |
| | Errors | Tokenizer, Orthographic spell checker |
| Character | Character types (letters, digits, etc.) | Character dictionary |
| | Character *n*-grams (fixed length) | – |
| | Character *n*-grams (variable length) | Feature selector |
| | Compression methods | Text compression tool |
| Syntactic | Part-of-speech (POS) | Tokenizer, Sentence splitter, POS tagger |
| | Chunks | Tokenizer, Sentence splitter, [POS tagger], Text chunker |
| | Sentence and phrase structure | Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser |
| | Rewrite rules frequencies | Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser |
| | Errors | Tokenizer, Sentence splitter, Syntactic spell checker |
| Semantic | Synonyms | Tokenizer, [POS tagger], Thesaurus |
| | Semantic dependencies | Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser |
| | Functional | Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries |
| Application-specific | Structural | HTML parser, Specialized parsers |
| | Content-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |
| | Language-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |

(Stamatatos, 2009)

# Type-token ratio

| | Types | Tokens | Corr. TTR |
|---|---|---|---|
| Eisenhover_1957 | 454 | 697 | 12.160 |
| Kennedy_1961 | 401 | 582 | 11.754 |
| Johnson_1965 | 373 | 550 | 11.246 |
| Nixon1_1969 | 547 | 840 | 13.345 |
| Nixon2_1973 | 364 | 674 | 9.914 |
| Carter_1977 | 365 | 487 | 11.695 |
| Reagan1_1981 | 648 | 942 | 14.929 |
| Reagan2_1985 | 684 | 1096 | 14.610 |
| Bushse_1989 | 546 | 881 | 13.007 |
| Clinton1_1993 | 445 | 661 | 12.239 |
| Clinton2_1997 | 548 | 884 | 13.033 |
| Bushju1_2001 | 440 | 670 | 12.020 |
| Bushju2_2005 | 577 | 909 | 13.533 |
| Obama_2009 | 721 | 977 | 16.311 |

Analiza inauguracijskih govora američkih predsjednika (www.tllab.it)

# Funkcijske riječi (stopwords)

always
am
amid
amidst
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are

i'm
immediate
in
inasmuch
inc
inc.
indeed
indicate
indicated
indicates
inner
inside
insofar
instead
into
inward
is
isn't
it
it'd
it'll
its

somebody
someday
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
take
taken

# Funkcijske riječi (stopwords)

a
ah
aha
aj
aja
ajme
ajooj
ajoooj
ako
akoli
alaj
ali
ama
amo
amo-tamo
ao
aoj
au
avaj
ba
bar
barem
baš
bez
bi
bih
bijah
bijahu
bijaše
bijasmo
bijaste
bijehu
bila
bile
bili
bilo
bilokako
bilokakva
bilošto
bio
bismo
će
čeg
čega
čem
ćemo
čemu
ćeš
često
ćete
četiri
četiriju
četirima
četiristo
četiristoti
četirma
četrdeset
četrdesetak
četrdeseterim
četrdeseterima
četrdesetero
četrdeseteroga
četrdeseterome
četrdeseteromu
četrdeseti
četrdesetoro
četrdesetoroga
četrdesetorome
četrdesetoromu
četri
četristotinjak
četrnaest
četrnaestak
četrnaesterim
četrnaesterima
četrnaestero
četrnaesteroga
četrnaesterome
četrnaesteromu
četrnaesti
četrnaestoro
četrnaestoroga
čiji
čijih
čijim
čijima
čijoj
čijom
čiju
čik
čim
čime
ću
da
dabome
dakako
dakle
danas
dapače
dašta
davno
de
ded
dede
deder
der
deset
deseta
desete
deseterim
deseterima
desetero
deseteroga
deseterome
deseteromu
deseti
desetih
desetim
desetima
deseto
desetog
desetoga
desetoj
deveti
devetih
devetim
devetima
devetnaest
devetnaesterim
devetnaesterima
devetnaestero
devetnaesteroga
devetnaesterome
devetnaesteromu
devetnaesti
devetnaestoro
devetnaestoroga
devetnaestorome
devetnaestoromu
deveto
devetog
devetoga
devetoj
devetom
devetome
devetoro
devetoroga
devetorome
devetoromu
devetsto
devetstoti
devetstotinjak
devetu
diljem
djelomice
djelomično
do
dobrano
doduše
dogodine
doista
dok
dokad
dokle
drugome
drugu
dum
duž
dva
dvadeset
dvadesetak
dvadeseterim
dvadeseterima
dvadesetero
dvadeseteroga
dvadeseterome
dvadeseteromu
dvadeseti
dvadesetoro
dvadesetorome
dvadesetoromu
dvaju
dvama
dvanaest
dvanaestak
dvanaesterim
dvanaesterima
dvanaestero
dvanaesteroga
dvanaesterome
dvanaesteromu
dvanaesti
dvanaestoro
dvanaestoroga
dvanaestorome
dvanaestoromu
dvaput
dve
dveju
dvema
dvije
dviju
dvjema
dvjesta
dvjesto

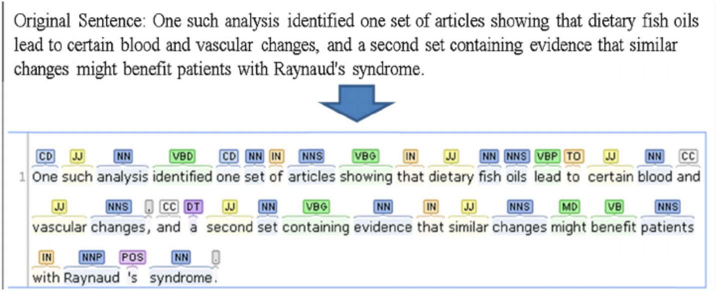
# N-grami

- N-grami riječi:

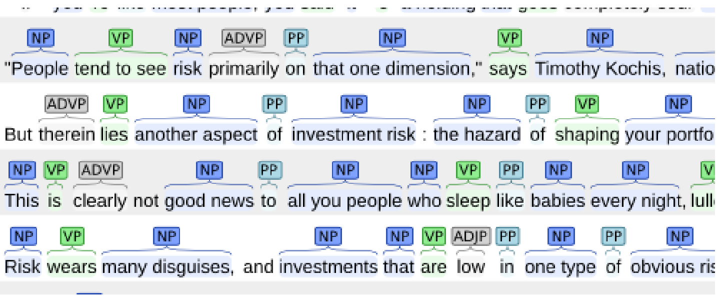
| | |
|---|---|
| Full sentence | It does not, however, control whether an exaction is within Congress's power to tax. |
| Unigrams | "It"; "does"; "not,"; "however,"; "control"; "whether"; "an"; "exaction"; "is"; "within"; "Congress's"; "power"; "to"; "tax." |
| Bigrams | "It does"; "does not,"; "not, however,"; "however, control"; "control whether"; "whether an"; "an exaction"; "exaction is"; "is within"; "within Congress's"; "Congress's power"; "power to"; "to tax." |
| Trigrams | "It does not"; "does not, however"; "not, however, control"; "however, control whether"; "control whether an"; "whether an exaction"; "an exaction is"; "exaction is within"; "is within Congress's"; "within Congress's power"; "Congress's power to"; "power to tax." |

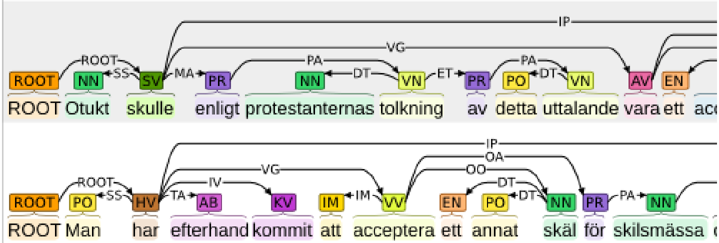
- N-grami slova:
  "Tko je to napisao?" $\Rightarrow$ Tko, ko_, o_j, _je, je_,...

# Sintaktičke značajke: POS tagging



# Sintaktičke značajke: parsanje



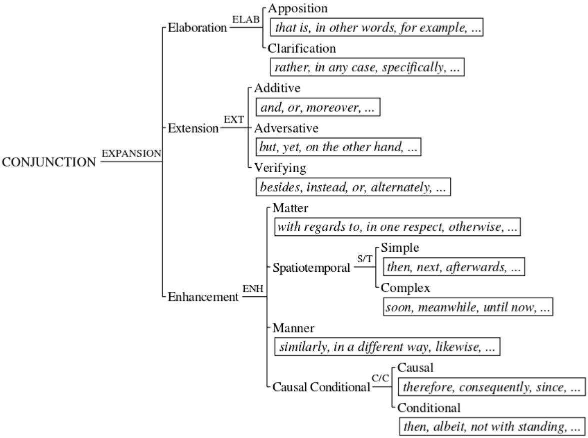
# Stilometrijske značajke

| Features | | Required tools and resources |
|---|---|---|
| Lexical | Token-based (word length, sentence length, etc.) | Tokenizer, [Sentence splitter] |
| | Vocabulary richness | Tokenizer |
| | Word frequencies | Tokenizer, [Stemmer, Lemmatizer] |
| | Word $n$-grams | Tokenizer |
| | Errors | Tokenizer, Orthographic spell checker |
| Character | Character types (letters, digits, etc.) | Character dictionary |
| | Character $n$-grams (fixed length) | – |
| | Character $n$-grams (variable length) | Feature selector |
| | Compression methods | Text compression tool |
| Syntactic | Part-of-speech (POS) | Tokenizer, Sentence splitter, POS tagger |
| | Chunks | Tokenizer, Sentence splitter, [POS tagger], Text chunker |
| | Sentence and phrase structure | Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser |
| | Rewrite rules frequencies | Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser |
| | Errors | Tokenizer, Sentence splitter, Syntactic spell checker |
| Semantic | Synonyms | Tokenizer, [POS tagger], Thesaurus |
| | Semantic dependencies | Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser |
| | Functional | Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries |
| Application-specific | Structural | HTML parser, Specialized parsers |
| | Content-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |
| | Language-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |

(Stamatatos, 2009)

# Sistemska funkcionalna gramatika

(Haliday, 1994)




(Argamon et al., 2007)

# Stilometrijske značajke

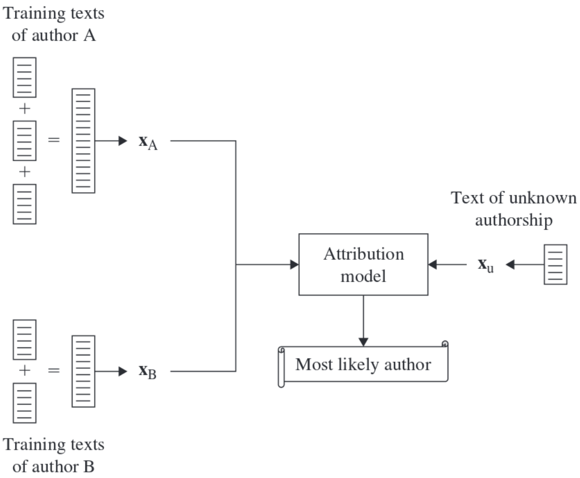
| Features | | Required tools and resources |
| --- | --- | --- |
| Lexical | Token-based (word length, sentence length, etc.) | Tokenizer, [Sentence splitter] |
| | Vocabulary richness | Tokenizer |
| | Word frequencies | Tokenizer, [Stemmer, Lemmatizer] |
| | Word $n$-grams | Tokenizer |
| | Errors | Tokenizer, Orthographic spell checker |
| Character | Character types (letters, digits, etc.) | Character dictionary |
| | Character $n$-grams (fixed length) | – |
| | Character $n$-grams (variable length) | Feature selector |
| | Compression methods | Text compression tool |
| Syntactic | Part-of-speech (POS) | Tokenizer, Sentence splitter, POS tagger |
| | Chunks | Tokenizer, Sentence splitter, [POS tagger], Text chunker |
| | Sentence and phrase structure | Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser |
| | Rewrite rules frequencies | Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser |
| | Errors | Tokenizer, Sentence splitter, Syntactic spell checker |
| Semantic | Synonyms | Tokenizer, [POS tagger], Thesaurus |
| | Semantic dependencies | Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser |
| | Functional | Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries |
| Application-specific | Structural | HTML parser, Specialized parsers |
| | Content-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |
| | Language-specific | Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries |

(Stamatatos, 2009)

# Metode

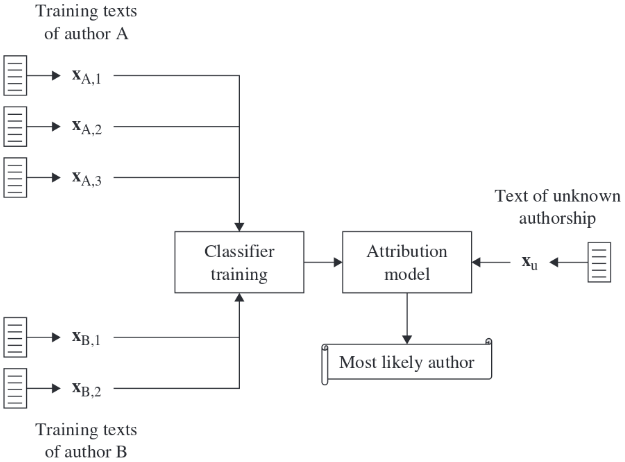
- Usporedba s **profilom** (starije metode)
  - probabilistički model $P(X|A)$
  - kompresija
  - zajednički n-grami
- Usporedba **primjerima** (nove metode)
  - vektorski model
  - sličnost: "Delta-metoda" (Burrows, 2002)
  - kompresija
  - demaskiranje

# Usporedba s profilom




(Stamatatos, 2009)

# Usporedba s primjerima




(Stamatatos, 2009)

# Atribucija autorstva vs. klasifikacija teksta

- Najčešće riječi (stopwords) su diskriminativne
- Ograničen skup za treniranje
- Neuravnotežena distribucija primjera

# Studija: Koppel et al. (2009)

- Podatci:
  - poruke e-pošte autora
  - po dvije knjige devetoro američkih i britanskih spisatelja (19./20. st.)
  - objave 20 mlađih blogera
- Pet algoritama strojnog učenja
- Stilističke i nestilističke (sadržajne) značajke

# Studija: Koppel et al. (2009)

| | |
|---|---|
| FW | A list of 512 function words, including conjunctions, prepositions, pronouns, modal verbs, determiners, and numbers (purely stylistic) |
| POS | Thirty-eight part-of-speech unigrams and 1,000 most common bigrams using the Brill (1992) part-of-speech tagger (purely stylistic) |
| SFL | All 372 nodes in SFL trees for conjunctions, prepositions, pronouns, and modal verbs (purely stylistic) |
| CW | The 1,000 words with highest information gain (Quinlan, 1986) in the training corpus among the 10,000 most common words in the corpus |
| CNG | The 1,000 character trigrams with highest information gain in the training corpus among the 10,000 most common trigrams in the corpus (cf. Keselj, 2003) |
| NB | WEKA's implementation (Witten & Frank, 2000) of Naïve Bayes (Lewis, 1998) with Laplace smoothing |
| J4.8 | WEKA's implementation of the J4.8 decision tree method (Quinlan, 1986) with no pruning |
| RMW | Our implementation of a version of Littlestone's (1988) Winnow algorithm, generalized to handle real-valued features and more than two classes (Schler, 2007) |
| BMR | Genkin et al.'s (2006) implementation of Bayesian multiclass regression |
| SMO | Weka's implementation of Platt's (1998) SMO algorithm for SVM with a linear kernel and default settings |

# Studija: Koppel et al. (2009)

## E-pošta

TABLE 2. Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the e-mail corpus.

| Features/learner | NB (%) | J4.8 (%) | RMW (%) | BMR (%) | SMO (%) |
|---|---|---|---|---|---|
| FW | 60.2 | 58.7 | 66.1 | 68.2 | 63.8 |
| POS | 61.0 | 59.0 | 66.1 | 66.3 | 67.1 |
| FW + POS | 65.9 | 61.6 | 68.0 | 67.8 | 71.7 |
| SFL | 57.2 | 57.2 | 65.6 | 67.2 | 62.7 |
| CW | 67.1 | 66.9 | 74.9 | 78.4 | 74.7 |
| CNG | 72.3 | 65.1 | 73.1 | 80.1 | 74.9 |
| CW + CNG | 73.2 | 68.9 | 74.2 | 83.6 | 78.2 |

# Studija: Koppel et al. (2009)

## Književnost

TABLE 3. Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the literature corpus.

| Features/learner | NB (%) | J4.8 (%) | RMW (%) | BMR (%) | SMO (%) |
|---|---|---|---|---|---|
| FW | 51.4 | 44.0 | 63.0 | 73.8 | 77.8 |
| POS | 45.9 | 50.3 | 53.3 | 69.6 | 75.5 |
| FW + POS | 56.5 | 46.2 | 61.7 | 75.0 | 79.5 |
| SFL | 66.1 | 45.7 | 62.8 | 76.6 | 79.0 |
| CW | 68.9 | 50.3 | 57.0 | 80.0 | 84.7 |
| CNG | 69.1 | 42.7 | 49.4 | 80.3 | 84.2 |
| CW + CNG | 73.9 | 49.9 | 57.1 | 82.8 | 86.3 |

# Studija: Koppel et al. (2009)

## Blogovi

TABLE 4. Accuracy test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the blog corpus.

| Features/learner | NB (%) | J4.8 (%) | RMW (%) | BMR (%) | SMO (%) |
|---|---|---|---|---|---|
| FW | 38.2 | 30.3 | 51.8 | 63.2 | 63.2 |
| POS | 34.0 | 30.3 | 51.0 | 63.2 | 60.6 |
| FW + POS | 47.0 | 34.3 | 62.3 | 70.3 | 72.0 |
| SFL | 35.4 | 36.3 | 61.4 | 69.2 | 71.7 |
| CW | 56.4 | 51.0 | 62.9 | 72.5 | 70.5 |
| CNG | 65.0 | 48.9 | 67.1 | 80.4 | 80.9 |
| CW + CNG | 69.9 | 51.6 | 75.4 | 86.1 | 85.7 |

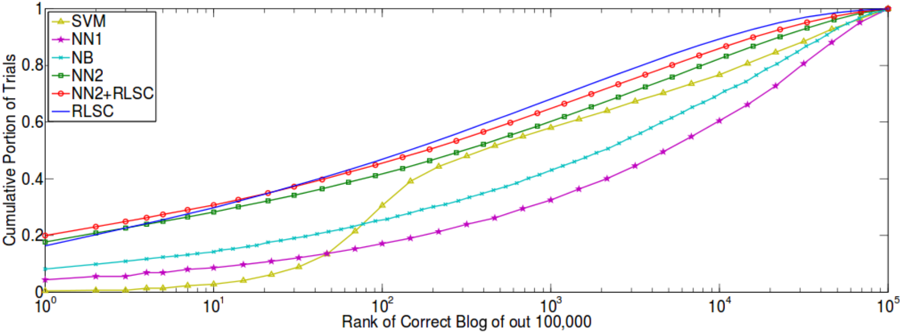
# Atribucija autorstva u big data

- Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012, May). **On the feasibility of internet-scale author identification**. In 2012 IEEE Symposium on Security and Privacy (pp. 300-314). IEEE.

# Atribucija autorstva u big data

- Problem **privatnosti**: anonimnost = privatnost
- 2.4 milijuna postova sa 100.000 blogova
- Eksperimenti sa nizom algoritama strojnog učenja
- Jednostavni modeli (k-nn) rade vrlo dobro
- Uzorak od 3 postova podudaran na postove ostalih autora (pomiješan sa 100.000 drugih blogova)
- Točan autor nalazi se u 20% slučajeva
- U 35% slučajeva, autor je u prvih 20 pogodaka

# Atribucija autorstva u big data




(Narayanan et al., 2012)

# Nedostatci studije

- Ograničenost na istu domenu
- Žrtva nije pokušala sakriti/izmijeniti svoj stil

# Skrivanje autorstva

- Brennan, M. R., & Greenstadt, R. (2009). **Practical Attacks Against Authorship Recognition Techniques**. In IAAI.

# Skrivanje autorstva

- Napad **skrivanjem** i napad **imitacijom**
- 15 sudionika:
  - autorski tekst (500 riječi)
  - skrivanje identiteta (500 riječi na zadanu temu)
  - imitacija: prepričati svoj dan u stilu Cormac McCarthya (roman "Cesta")
- Zaključak: sve se metode mogu vrlo lako zavarati

# Skrivanje autorstva

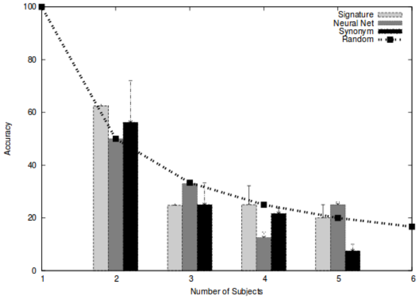



Figure 2: Accuracy in detecting obfuscation attacks. The x-axis shows the number of subjects, the y-axis shows the average percentage of obfuscation attacks correctly classified. The error bars show the standard error for each experiment.

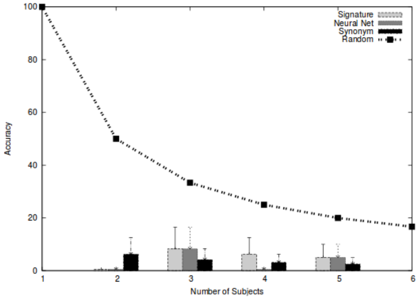



Figure 3: Accuracy in detecting imitation attacks. The x-axis shows the number of subjects, the y-axis shows the average percentage of imitation attacks correctly classified. The error bars show the standard error for each experiment.

(Brennan, M. R., & Greenstadt, R., 2009)

# Vođeno skrivanje autorstva

- Kacmarcik, G., & Gamon, M. (2006). **Obfuscating document stylometry to preserve author anonymity**. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 444-451). Association for Computational Linguistics.

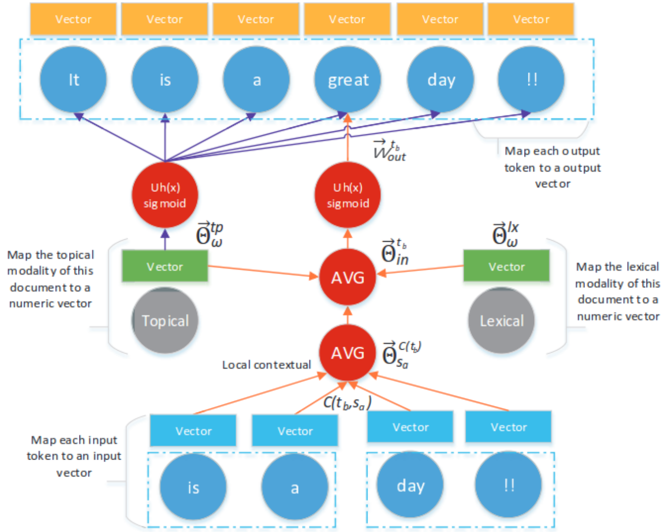
# Vođeno skrivanje autorstva

- Koliko je lako autoru prezentirati potrebne izmjene?
- Koliko su postojeće metode otporne na ovakve izmjene?
- Koliko je rada potrebno uložiti u skrivanje?

# Učenje stilometrijske reprezentacije

- Ding, S. H., Fung, B., Iqbal, F., & Cheung, W. K. (2016). **Learning Stylometric Representations for Authorship Analysis**. arXiv preprint arXiv:1606.01219.

# Učenje stilometrijske reprezentacije




(Ding et al., 2016)

# Plan

# Provjera autorstva

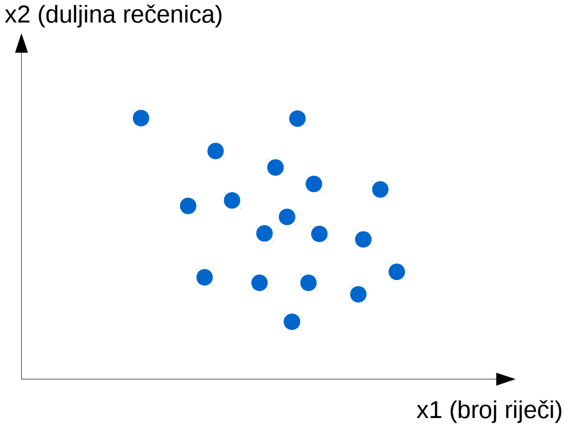
- Imamo primjere teksta jednoga autora, trebamo identificirati je li text X pisao isti taj autor
- Ne postoji popis mogućih autora!
- Teži problem od atribucije autorstva: ne postoji puno radova!
- Problem **negativnih primjera**
    - što je reprezentativan uzorak ne-Shakespearovih tekstova?

# Provjera autorstva

- Naivan pristup:
  - uzorkovati reprezentativnu zbirku tekstova čiji autor nije A
  - trenirati **binarni klasifikator** A vs. ne-A
  - konceptualni problem: novi tekst nekog novog autora može biti sličniji A nego ne-A
- Bolji pristupi:
  - **jednoklasna klasifikacija**
  - jesu li tekstovi X i Y nastali od istog autora? $\Rightarrow$ **demaskiranje**

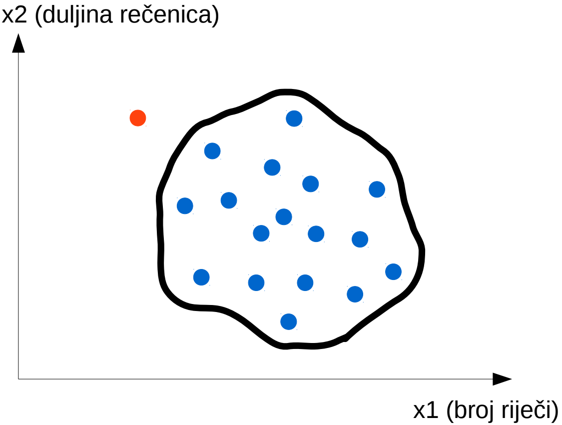
# Jednoklasni klasifikator

## One-class SVM



# Jednoklasni klasifikator

## One-class SVM



# Demaskiranje (Koppel et al.,2009)

- Nathaniel Hawthorne:
  "Kuća sa sedam zabata" vs. "Grimizno slovo"
- Izražene, ali ograničene razlike ("he" vs. "she")
- Ideja: stilističke razlike između tekstova istog autora su manje od razlika između tekstova različitih autora
- Iterativno eliminirati značajki klasifikatora
- Tekstovi koje klasifikator ne uspijeva više razlikovati tekstovi su **istog autora**
- Tekstovi različitih autora imaju više različitosti, pa ih klasifikator i dalje uspješno razlikuje

# Demaskiranje (Koppel et al.,2009)

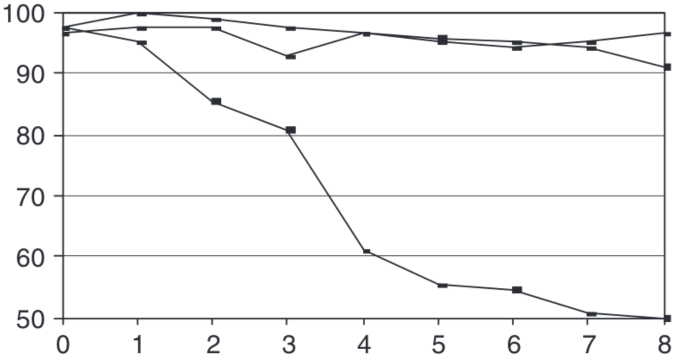



FIG. 3. Tenfold cross-validation accuracy of models distinguishing *House of Seven Gables* from each of Hawthorne, Melville, and Cooper. The *x* axis represents the number of iterations of eliminating best features at previous iteration. The curve well below the others is that of Hawthorne, the actual author.

# Plan

1. NLP i strojno učenje
2. Atribucija autorstva
3. Provjera autorstva
4. Profiliranje autora

# Profiliranje autora

- Imamo tekst anonimnog autora, nemamo kadnidate, želimo zaključiti o karakteristikama autora
- Sociolingvistika: **različite grupe ljudi** jezik koriste na **različit način**
- Identične metode kao i za atribuciju autorstva, ali ih primijenjujemo kako bismo razlikovali **grupe autora**, a ne pojedinačne autore
- **Demografske značajke**: spol, dob, nacionalnost, etnička pripadnost, materinji jezik, politička orijentacija, preference prema brendovima, bračni status, prihod, velepetori model ličnosti

# Velepetori model ličnosti

| Trait | Description |
|---|---|
| Openness | Curious, original, intellectual, creative, and open to new ideas. |
| Conscientiousness | Organized, systematic, punctual, achievement oriented, and dependable. |
| Extraversion | Outgoing, talkative, sociable, and enjoys being in social situations. |
| Agreeableness | Affable, tolerant, sensitive, trusting, kind, and warm. |
| Neuroticism | Anxious, irritable, temperamental, and moody. |

http://www.web-books.com/eLibrary/ON/B0/B58/010MB58.html

# Studija: Koppel et al. (2009)

- **Spol+dob**: 47.000 blogova s informacijama koje su dali autori
- **Materinji jezik**: International Corpus of Learner English (L2)
- **Osobine ličnosti**: neurotičnost
  - 20-minutni eseji studenata u stilu toka svijesti
  - upitnik za peterofaktorski model

# Studija: Koppel et al. (2009)

TABLE 5. Classification accuracy for profiling problems using different feature sets.

| | Baseline | Style | Content | Style + Content |
|---|---|---|---|---|
| Gender (2 classes) | *50.0* | 72.0 | 75.1 | **76.1** |
| Age (3 classes) | *42.7* | 66.9 | 75.5 | **77.7** |
| Language (5 classes) | *20.0* | 65.1 | **82.3** | 79.3 |
| Neuroticism (2 classes) | *50.0* | **65.7** | 53.0 | 63.1 |

# Studija: Koppel et al. (2009)

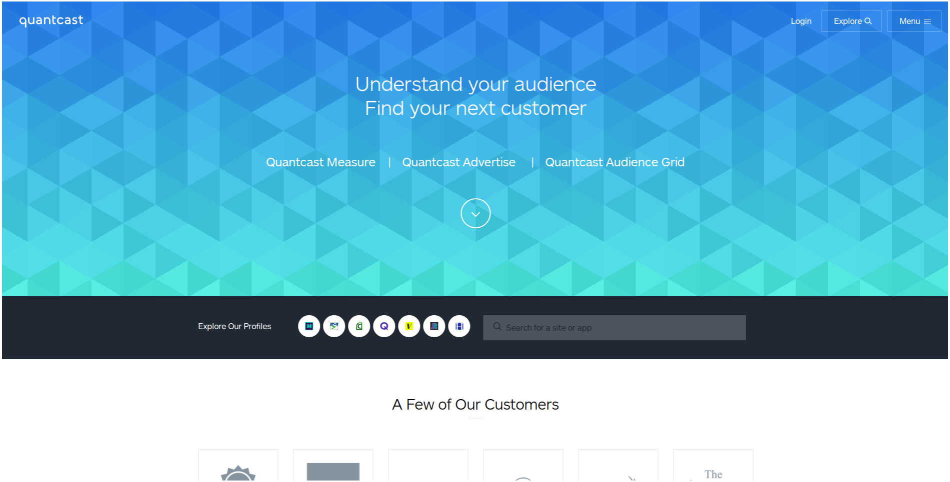
TABLE 6. Most important style and content features (by information gain) for each class of texts in each profiling problem.

| Class | Style features | Content features |
|---|---|---|
| Female | **personal pronoun,** *I, me, him, my* | *cute, love, boyfriend, mom, feel* |
| Male | **determiner,** *the, of,* **preposition-matter,** *as* | *system, software, game, based, site* |
| Teens | *im, so, thats, dont, cant* | *haha, school, lol, wanna, bored* |
| 20s | **preposition, determiner,** *of, the, in* | *apartment, office, work, job, bar* |
| 30s+ | **preposition,** *the,* **determiner,** *of, in* | *years, wife, husband, daughter, children* |
| Bulgarian | **conjunction-extension, pronoun-interactant,** *however,* **pronoun-conscious,** *and* | *bulgaria, university, imagination, bulgarian, theoretical* |
| Czech | **personal pronoun,** *usually, did, not, very* | *czech, republic, able, care, started* |
| French | *indeed,* **conjunction-elaboration,** *will,* **auxverb-future, auxverb-probability** | *identity, europe, european, nation, gap* |
| Russian | *can't,* *i, can, over, every* | *russia, russian, crimes, moscow, crime* |
| Spanish | **determiner-specific,** *this, going_to, because, although* | *spain, restoration, comedy, related, hardcastle* |
| Neurotic | *myself,* **subject pronoun, reflexive pronoun, preposition-behalf, pronoun-speaker** | *put, feel, worry, says, hurt* |
| Non-neurotic | *little,* **auxverbs-obligation, nonspecific determiner,** *up,* **preposition-agent** | *reading, next, cool, tired, bed* |

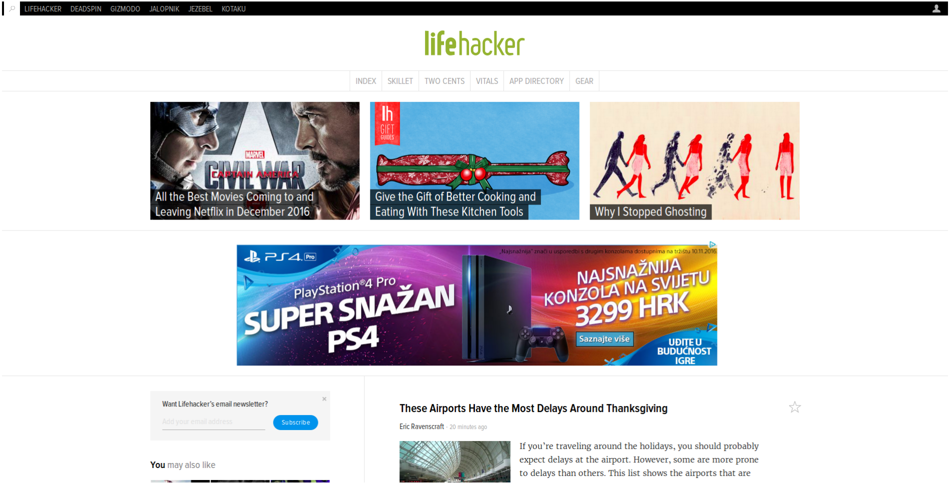
# Profiliranje korisnika Twittera

- Culotta, A., Ravi, N. K., & Cutler, J. (2016). **Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data**. Journal of Artificial Intelligence Research, 55, 389-408.

# Profiliranje korisnika Twittera



# Profiliranje korisnika Twittera



# Profiliranje korisnika Twittera

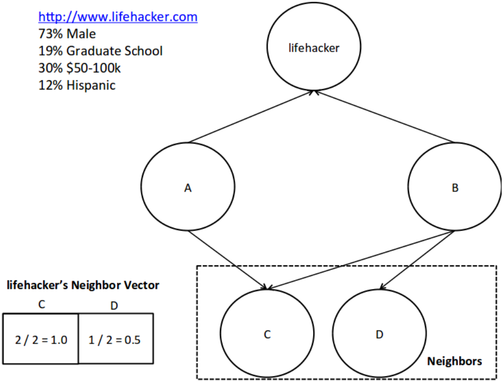


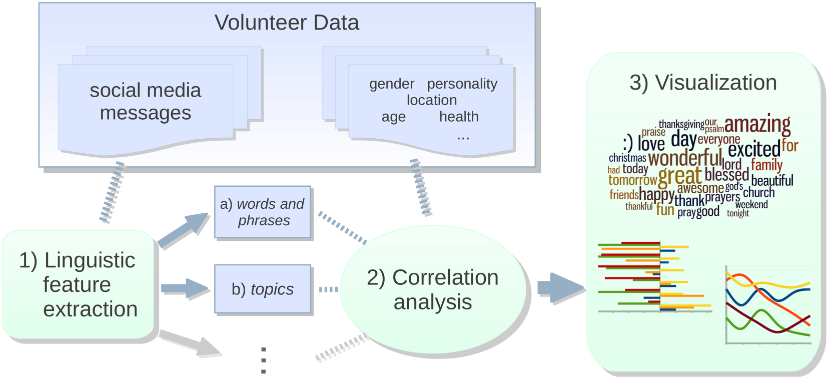

Figure 1: Data model. We collect QuantCast demographic data for each website, then construct a **Neighbor Vector** from the Twitter connections of that website, based on the proportion of the website's followers that are friends with each neighbor.

(Cullota et al., 2016)

# Jezik društvenih medija

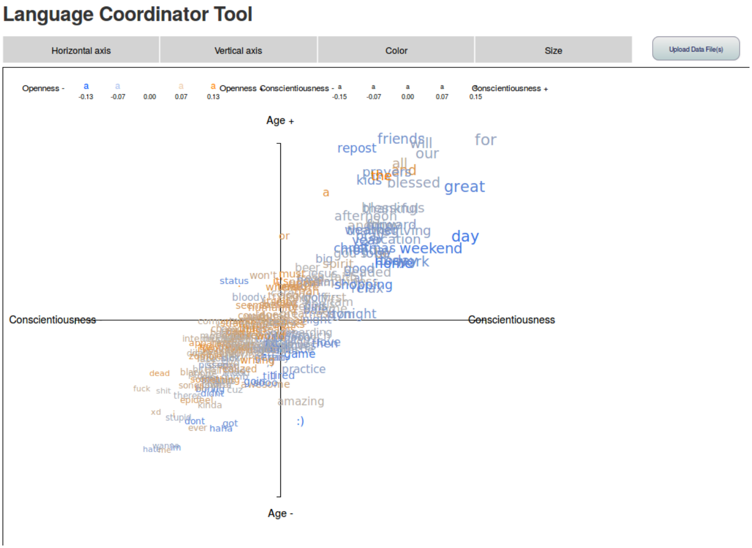
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., & Ungar, L. H. (2013). **Personality, gender, and age in the language of social media: The open-vocabulary approach**. PloS one, 8(9), e73791.

# Jezik društvenih medija




(Schwartz et al., 2013)

# Jezik društvenih medija




http://lexhub.org/langCoordinator/langCoordTool.html

# Up/downspeak


- Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011). **Extracting social power relationships from natural language**. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 773-782). Association for Computational Linguistics.

# Up/downspeak

## On Enron Emails

| Lect | Ngram | Example |
|---|---|---|
| UpSpeak | if you | "Let me know *if you* need anything." |
| | | "Please call me *if you* have any questions." |
| Down-Speak | give me | "Read this over and *give me* a call." |
| | | "Please *give me* your comments next week." |

| Lect | Ngram | Example |
|---|---|---|
| UpSpeak | I'll, we'll | "*I'll* let you know the final results soon" |
| | | "Everyone is very excited […] and we're confident *we'll* be successful" |
| DownSpeak | that is, this is | "Neither does any other group but *that is* not my problem" |
| | | "I think *this is* an excellent letter" |

(Bramsen et al., 2011)

# Plan

1. NLP i strojno učenje
2. Atribucija autorstva
3. Provjera autorstva
4. Profiliranje autora

# Otvoreni izazovi

- Problem duljine teksta
- Kako razlikovati između autorstva, žanra i teme
- Problem nedovoljne točnosti (za pravosuđe)
- Otvoreni skup autora
- Robusnost kroz teme i žanrove

# Perspektive

- Natjecanja PAN (godišnje, od 2007)
  - http://pan.webis.de/
- Sve veći interes za NLP u sociolingvistici
  - Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). **Computational sociolinguistics: A survey**. arXiv preprint arXiv:1508.07544.

Hvala na pažnji!

jan.snajder@fer.hr

TakeLab

takelab.fer.hr