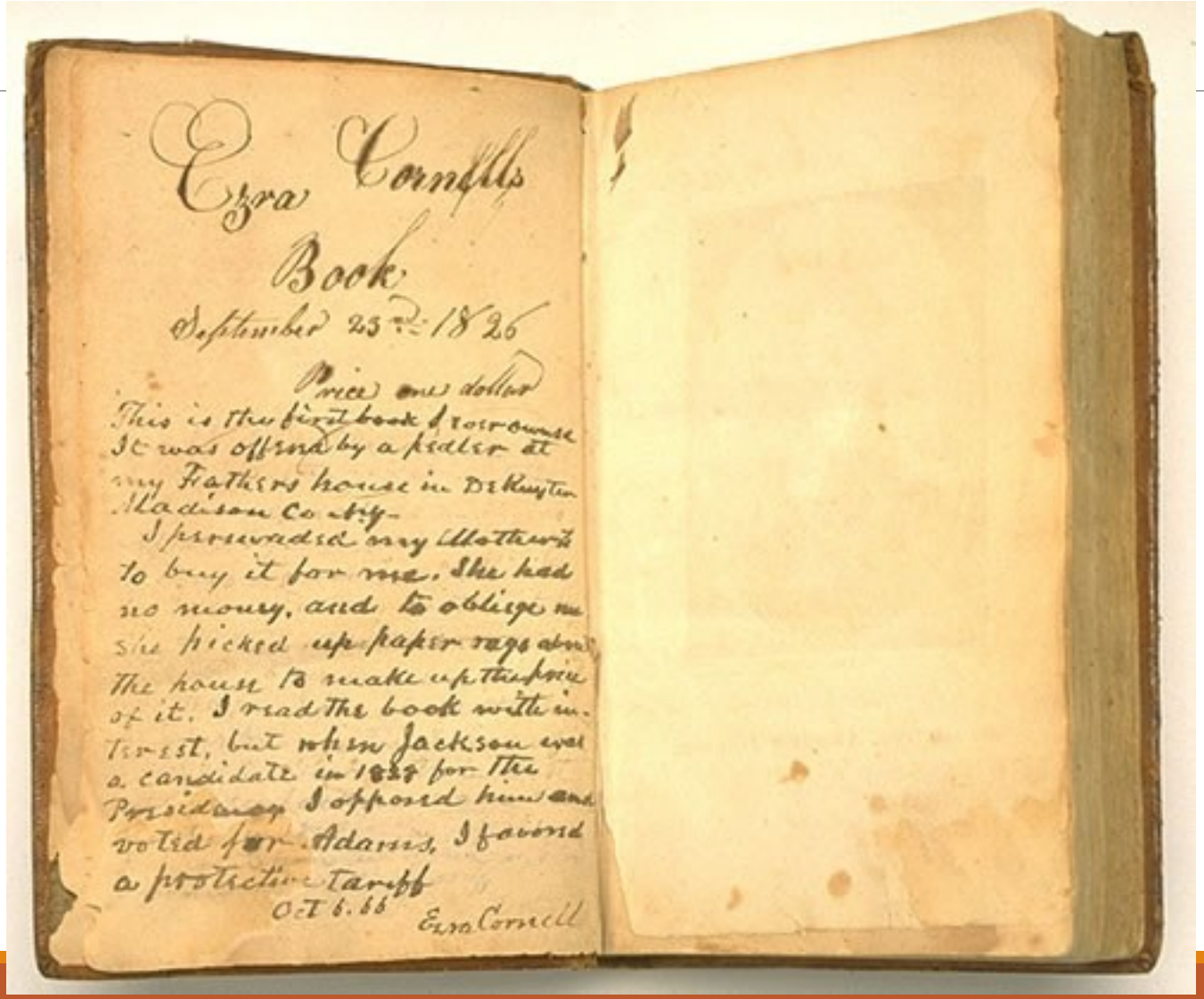
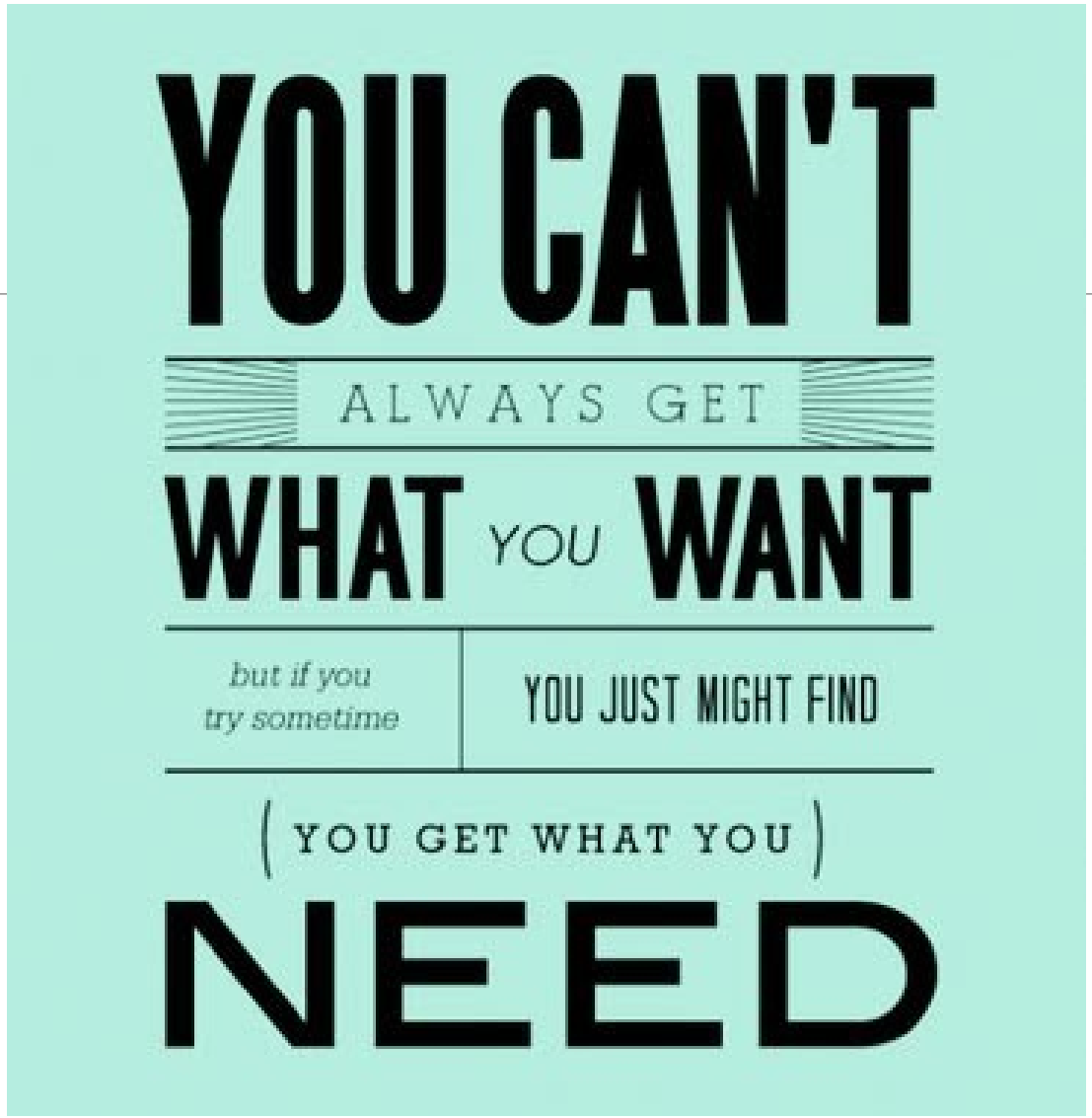
Základy kvantitativní analýzy textových dat

MARTIN BOROŠ

20.11.2015

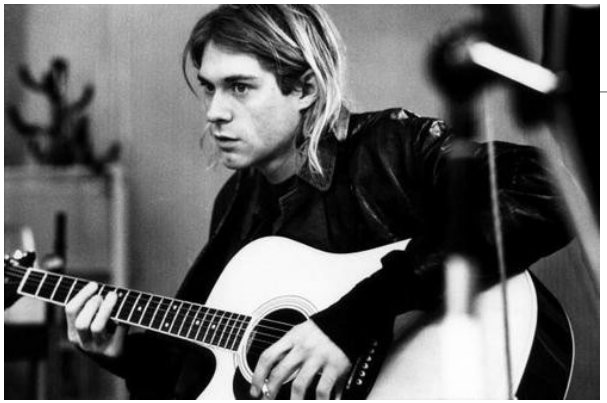
(Ne)Triviálny úvod

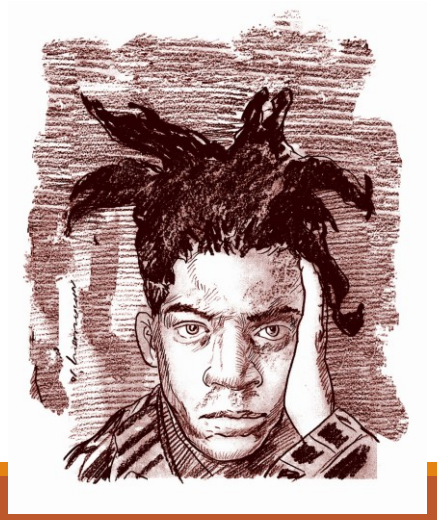



YOU CAN'T
ALWAYS GET
WHAT YOU WANT
but if you try sometime
YOU JUST MIGHT FIND
YOU GET WHAT YOU
NEED

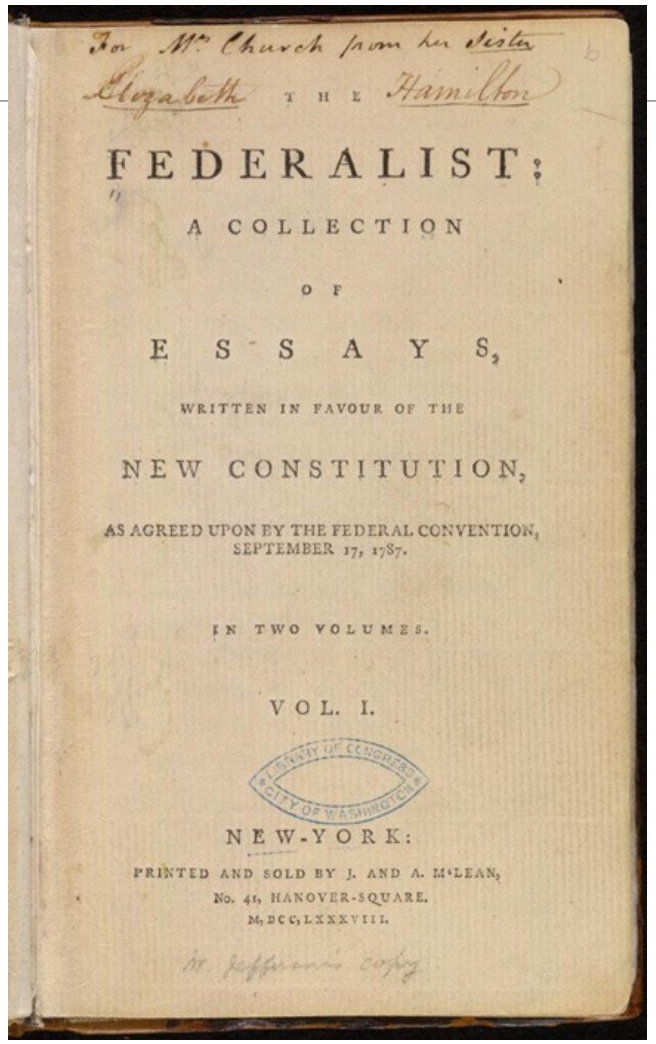
Klub 27?

Klub 27



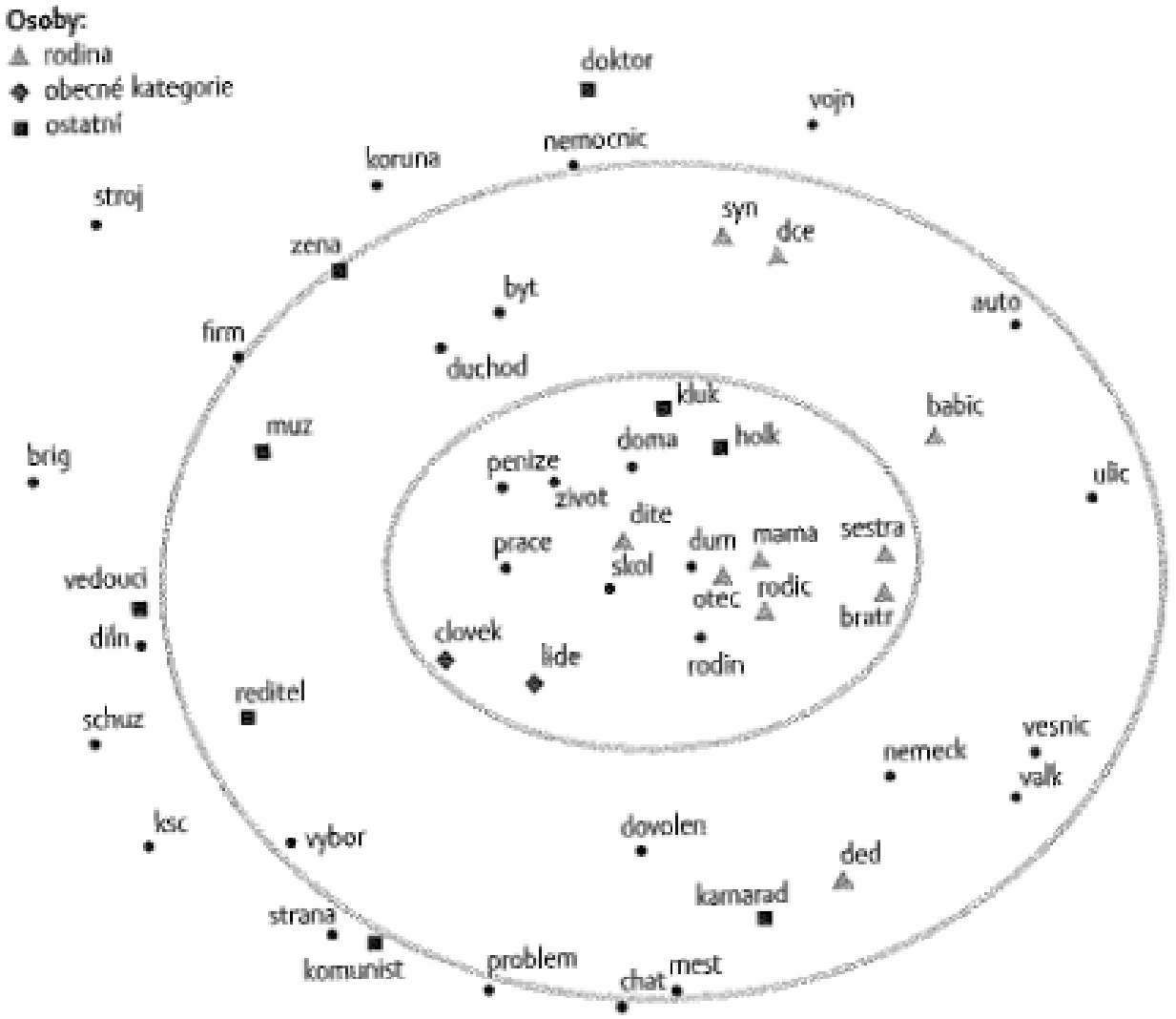


Tool RapidMiner



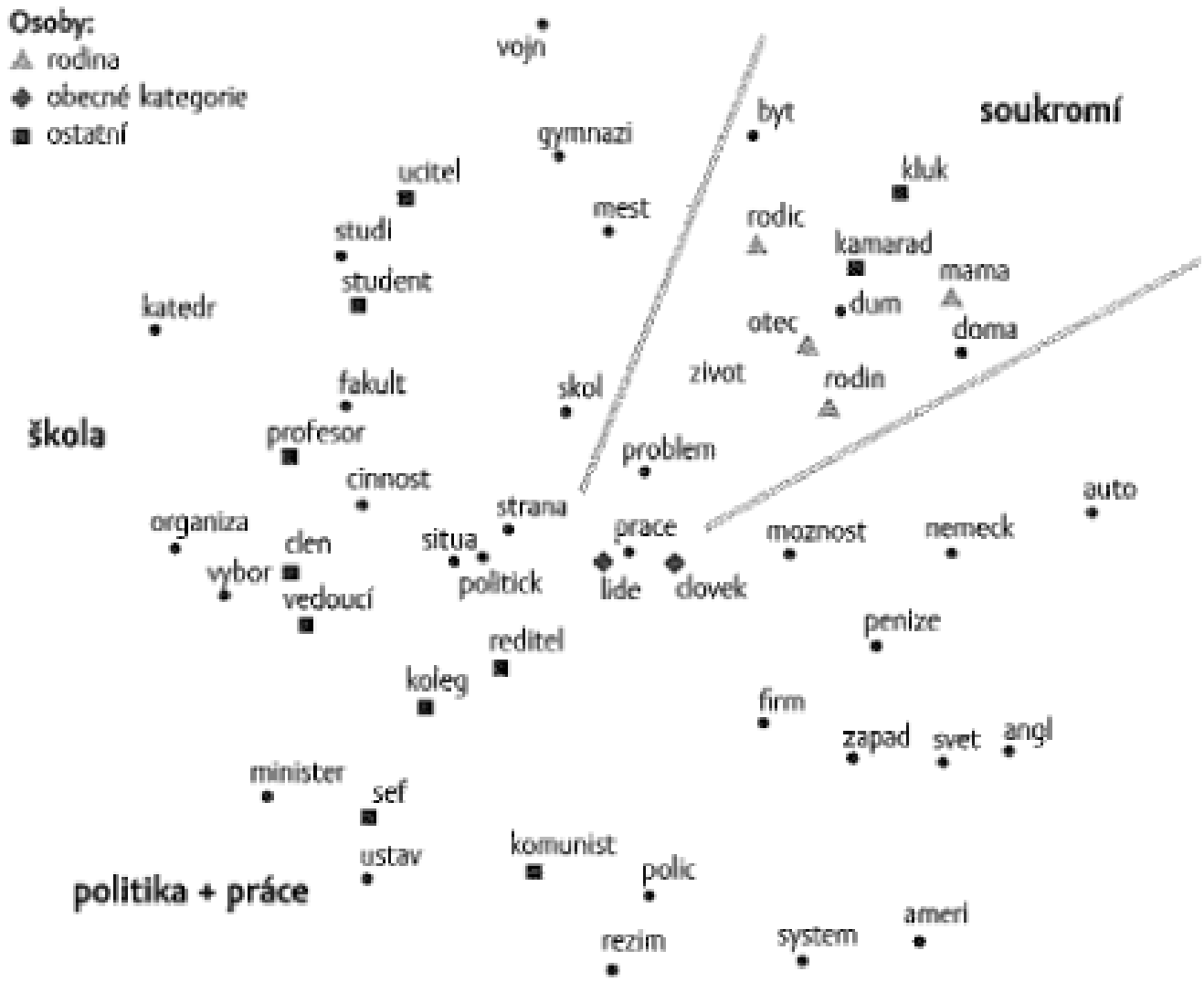
voyant-tools.org

Obrázek 1 Graf spoluvýskytů slov v biografických interview dělnic. Barevně jsou odlišena slova označující sociální role.



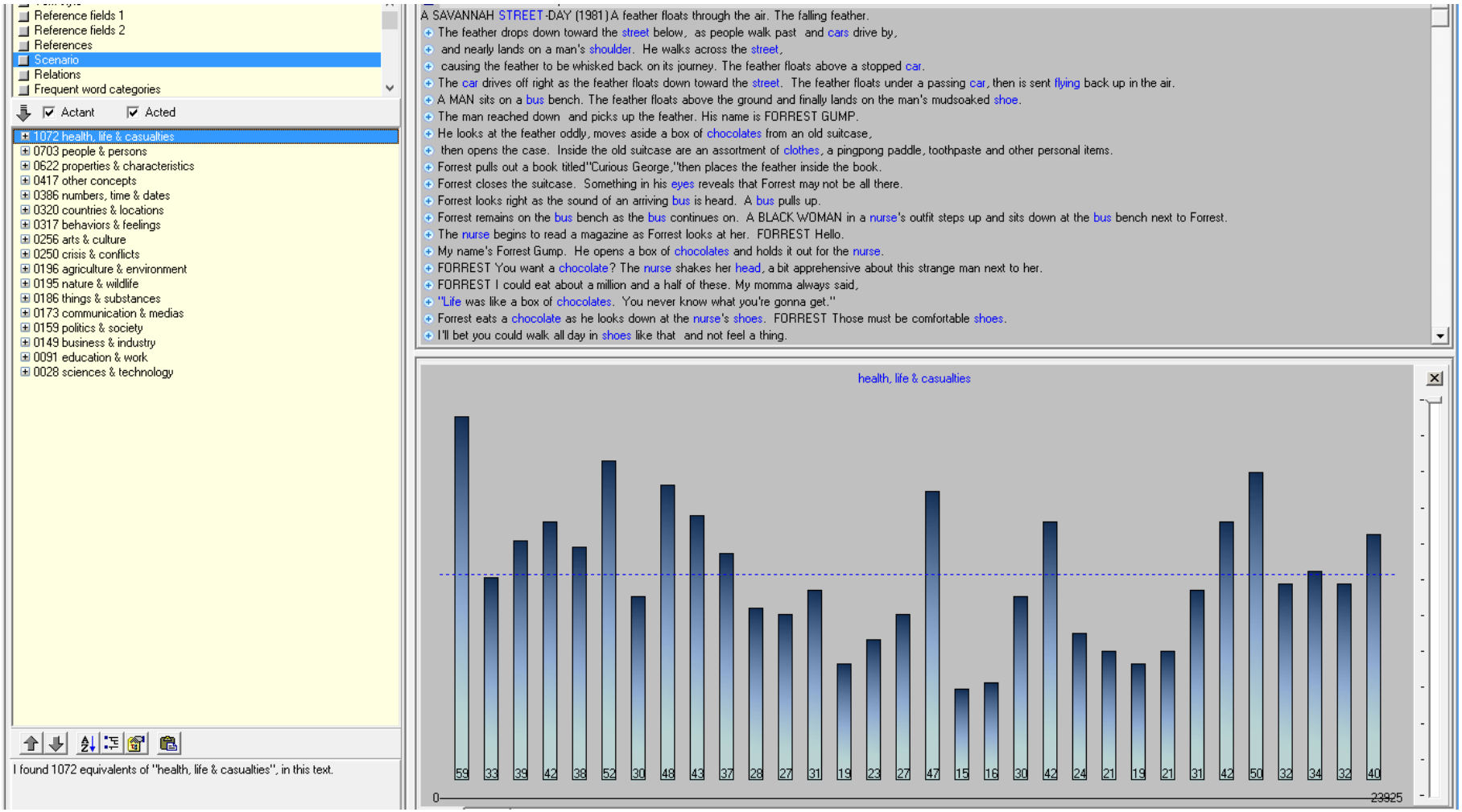
Hájek, M 2014 *Čtenář a stroj : vybrané metody sociálněvědní analýzy textů.* Praha: SLON.

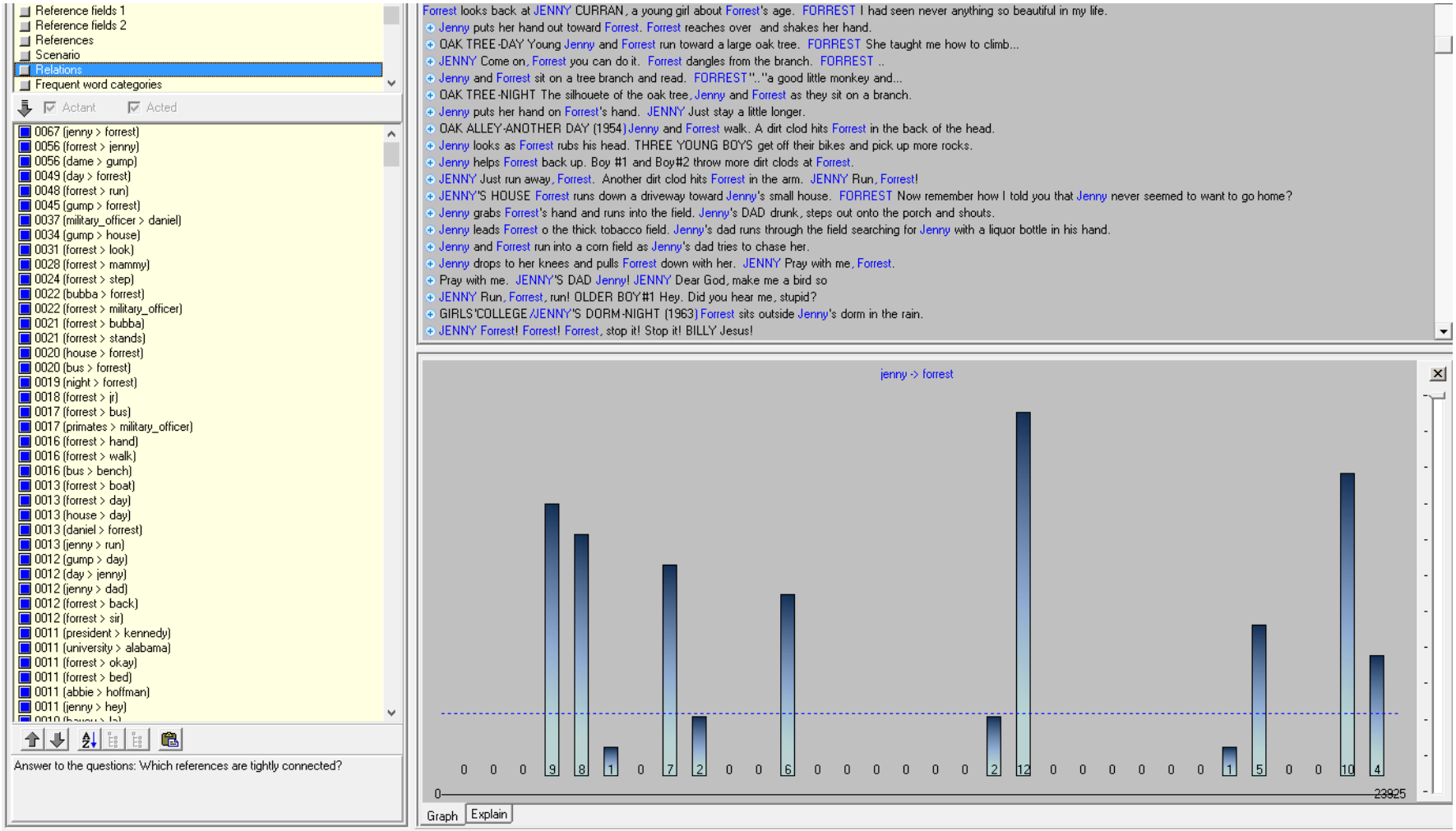
Obrázek 2 Graf spoluvýskytů slov v biografických interview příslušníků inteligence (muži). Barevně jsou odlišena slova označující sociální role.

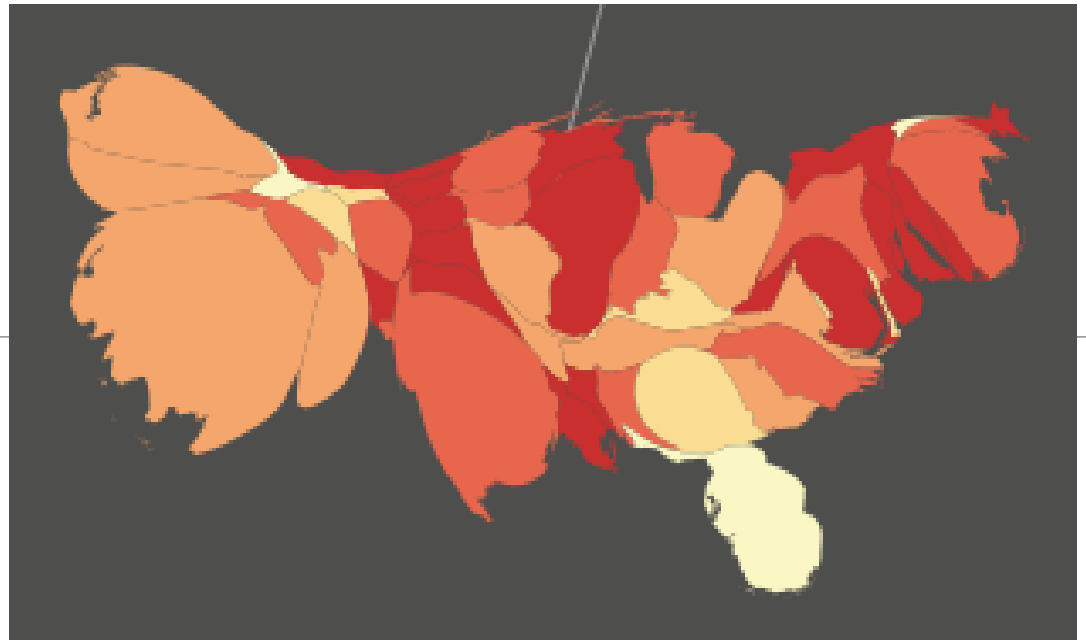


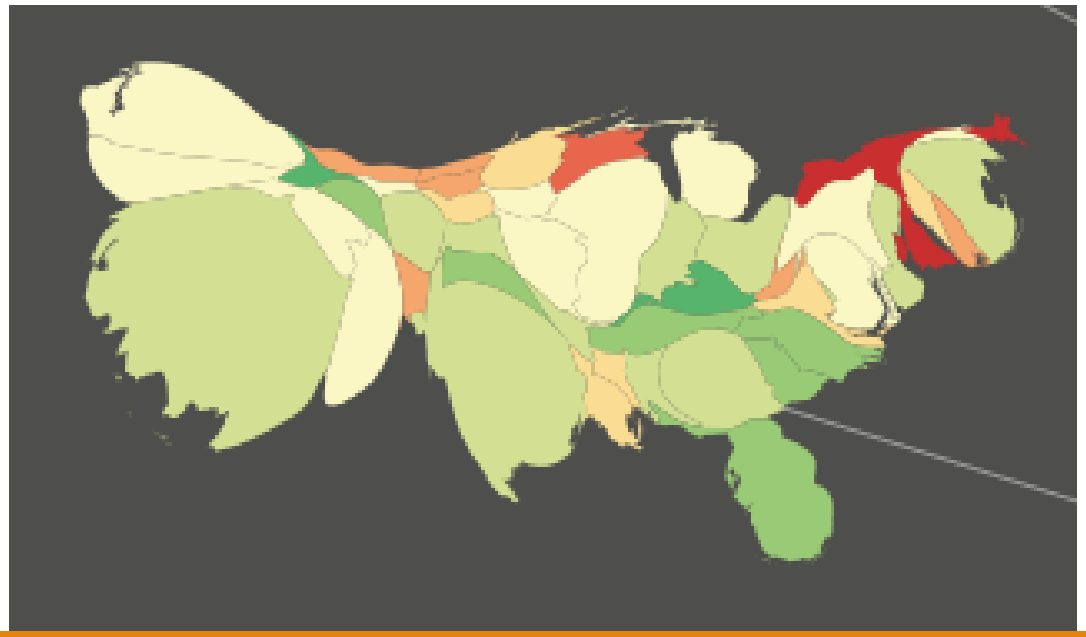
Hájek, M 2014 *Čtenář a stroj : vybrané metody sociálněvědní analýzy textů.* Praha: SLON.

Tool Tropes


Reference fields 1
Reference fields 2
References
Scenario
Relations
Frequent word categories
Actant
Acted
1072 health, life & casualties
0703 people & persons
0622 properties & characteristics
0417 other concepts
0386 numbers, time & dates
0320 countries & locations
0317 behaviors & feelings
0256 arts & culture
0250 crisis & conflicts
0196 agriculture & environment
0195 nature & wildlife
0186 things & substances
0173 communication & medias
0159 politics & society
0149 business & industry
0091 education & work
0028 sciences & technology
I found 1072 equivalents of "health, life & casualties", in this text.
A SAVANNAH STREET-DAY (1981) A feather floats through the air. The falling feather.
The feather drops down toward the street below, as people walk past and cars drive by,
and nearly lands on a man's shoulder. He walks across the street,
causing the feather to be whisked back on its journey. The feather floats above a stopped car.
The car drives off right as the feather floats down toward the street. The feather floats under a passing car, then is sent flying back up in the air.
A MAN sits on a bus bench. The feather floats above the ground and finally lands on the man's mudsoaked shoe.
The man reached down and picks up the feather. His name is FORREST GUMP.
He looks at the feather oddly, moves aside a box of chocolates from an old suitcase,
then opens the case. Inside the old suitcase are an assortment of clothes, a pingpong paddle, toothpaste and other personal items.
Forrest pulls out a book titled "Curious George," then places the feather inside the book.
Forrest closes the suitcase. Something in his eyes reveals that Forrest may not be all there.
Forrest looks right as the sound of an arriving bus is heard. A bus pulls up.
Forrest remains on the bus bench as the bus continues on. A BLACK WOMAN in a nurse's outfit steps up and sits down at the bus bench next to Forrest.
The nurse begins to read a magazine as Forrest looks at her. FORREST Hello.
My name's Forrest Gump. He opens a box of chocolates and holds it out for the nurse.
FORREST You want a chocolate? The nurse shakes her head, a bit apprehensive about this strange man next to her.
FORREST I could eat about a million and a half of these. My momma always said,
"Life was like a box of chocolates. You never know what you're gonna get."
Forrest eats a chocolate as he looks down at the nurse's shoes. FORREST Those must be comfortable shoes.
I'll bet you could walk all day in shoes like that and not feel a thing.
health, life & casualties
59 33 39 42 38 52 30 48 43 37 28 27 31 19 23 27 47 15 16 30 42 24 21 19 21 31 42 50 32 34 32 40
0
23925

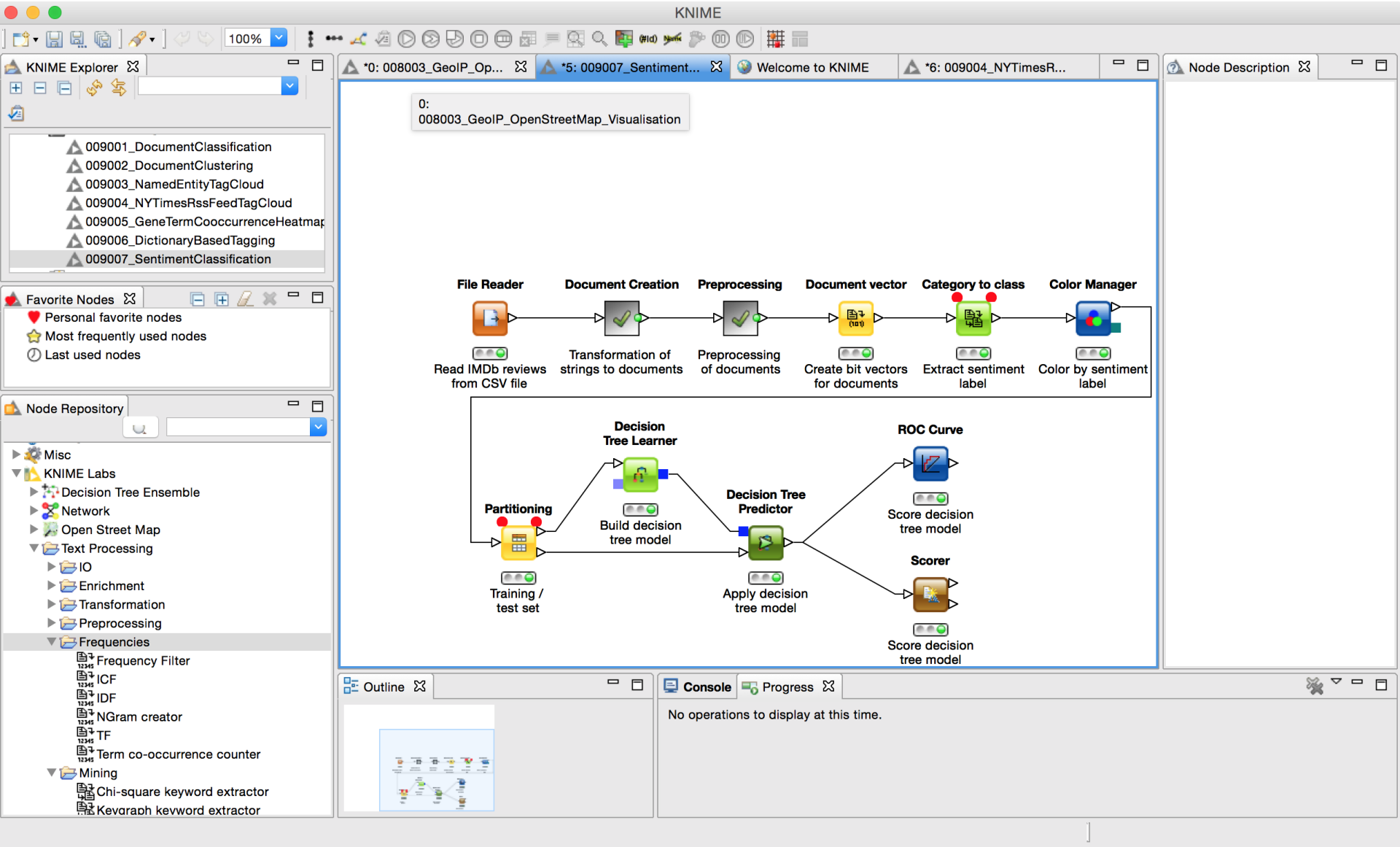

Reference fields 1
Reference fields 2
References
Scenario
Relations
Frequent word categories
Actant
Acted
0067 (jenny > forrest)
0056 (forrest > jenny)
0056 (dame > gump)
0049 (day > forrest)
0048 (forrest > run)
0045 (gump > forrest)
0037 (military_officer > daniel)
0034 (gump > house)
0031 (forrest > look)
0028 (forrest > mammy)
0024 (forrest > step)
0022 (bubba > forrest)
0022 (forrest > military_officer)
0021 (forrest > bubba)
0021 (forrest > stands)
0020 (house > forrest)
0020 (bus > forrest)
0019 (night > forrest)
0018 (forrest > jr)
0017 (forrest > bus)
0017 (primates > military_officer)
0016 (forrest > hand)
0016 (forrest > walk)
0016 (bus > bench)
0013 (forrest > boat)
0013 (forrest > day)
0013 (house > day)
0013 (daniel > forrest)
0013 (jenny > run)
0012 (gump > day)
0012 (day > jenny)
0012 (jenny > dad)
0012 (forrest > back)
0012 (forrest > sir)
0011 (president > kennedy)
0011 (university > alabama)
0011 (forrest > okay)
0011 (forrest > bed)
0011 (abbie > hoffman)
0011 (jenny > hey)
Answer to the questions: Which references are tightly connected?
Forrest looks back at JENNY CURRAN, a young girl about Forrest's age. FORREST I had seen never anything so beautiful in my life.
Jenny puts her hand out toward Forrest. Forrest reaches over and shakes her hand.
OAK TREE-DAY Young Jenny and Forrest run toward a large oak tree. FORREST She taught me how to climb...
JENNY Come on, Forrest you can do it. Forrest dangles from the branch. FORREST ..
Jenny and Forrest sit on a tree branch and read. FORREST "..."a good little monkey and...
OAK TREE-NIGHT The silhouete of the oak tree, Jenny and Forrest as they sit on a branch.
Jenny puts her hand on Forrest's hand. JENNY Just stay a little longer.
OAK ALLEY-ANOTHER DAY (1954) Jenny and Forrest walk. A dirt clod hits Forrest in the back of the head.
Jenny looks as Forrest rubs his head. THREE YOUNG BOYS get off their bikes and pick up more rocks.
Jenny helps Forrest back up. Boy #1 and Boy#2 throw more dirt clods at Forrest.
JENNY Just run away, Forrest. Another dirt clod hits Forrest in the arm. JENNY Run, Forrest!
JENNY'S HOUSE Forrest runs down a driveway toward Jenny's small house. FORREST Now remember how I told you that Jenny never seemed to want to go home?
Jenny grabs Forrest's hand and runs into the field. Jenny's DAD drunk, steps out onto the porch and shouts.
Jenny leads Forrest o the thick tobacco field. Jenny's dad runs through the field searching for Jenny with a liquor bottle in his hand.
Jenny and Forrest run into a corn field as Jenny's dad tries to chase her.
Jenny drops to her knees and pulls Forrest down with her. JENNY Pray with me, Forrest.
Pray with me. JENNY'S DAD Jenny! JENNY Dear God, make me a bird so
JENNY Run, Forrest, run! OLDER BOY#1 Hey. Did you hear me, stupid?
GIRLS'COLLEGE /JENNY'S DORM-NIGHT (1963) Forrest sits outside Jenny's dorm in the rain.
JENNY Forrest! Forrest! Forrest, stop it! Stop it! BILLY Jesus!
jenny -> forrest
0 0 0 9 8 1 0 7 2 0 0 6 0 0 0 0 0 0 2 12 0 0 0 0 0 0 1 5 0 0 10 4
0
23925
Graph
Explain

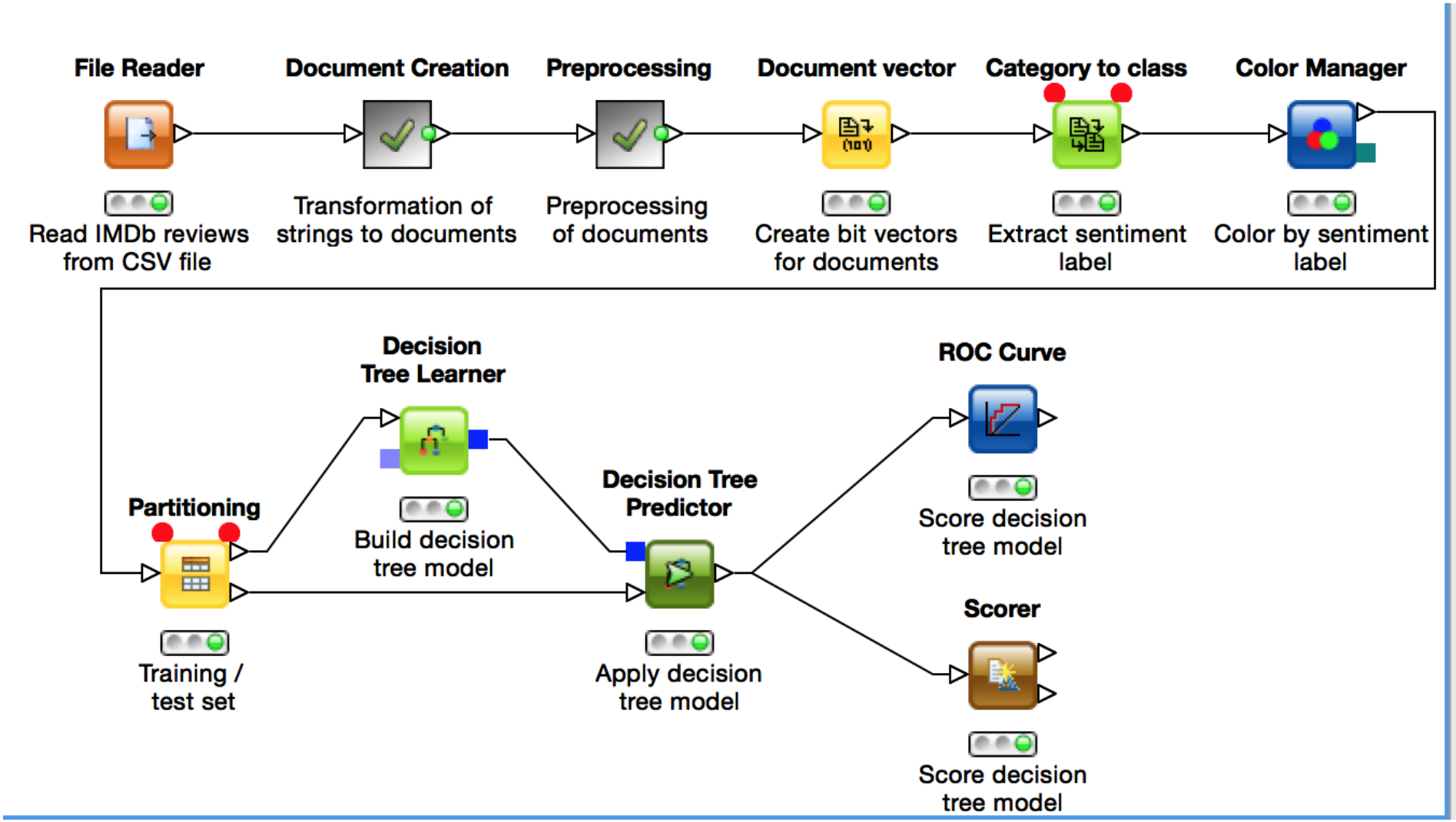




Mapa nálad podľa štátov USA o 15.00 a 22.00, podľa Mislove et al. 2010 Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter [online]. (cit. 18.11.2015). Dostupné na: http://www.ccs.neu.edu/home/amislove/twittermood

Tool KNIME


KNIME
KNIME Explorer
009001_DocumentClassification
009002_DocumentClustering
009003_NamedEntityTagCloud
009004_NYTimesRssFeedTagCloud
009005_GeneTermCooccurrenceHeatmap
009006_DictionaryBasedTagging
009007_SentimentClassification
Favorite Nodes
Personal favorite nodes
Most frequently used nodes
Last used nodes
Node Repository
Misc
KNIME Labs
Decision Tree Ensemble
Network
Open Street Map
Text Processing
IO
Enrichment
Transformation
Preprocessing
Frequencies
Frequency Filter
ICF
IDF
NGram creator
TF
Term co-occurrence counter
Mining
Chi-square keyword extractor
Keygraph keyword extractor
*0: 008003_GeoIP_Op...
*5: 009007_Sentiment...
Welcome to KNIME
*6: 009004_NYTimesR...
0:
008003_GeoIP_OpenStreetMap_Visualisation
File Reader
Document Creation
Preprocessing
Document vector
Category to class
Color Manager
Read IMDb reviews from CSV file
Transformation of strings to documents
Preprocessing of documents
Create bit vectors for documents
Extract sentiment label
Color by sentiment label
Decision Tree Learner
ROC Curve
Partitioning
Build decision tree model
Decision Tree Predictor
Score decision tree model
Scorer
Training / test set
Apply decision tree model
Score decision tree model
Node Description
Outline
Console
Progress
No operations to display at this time.

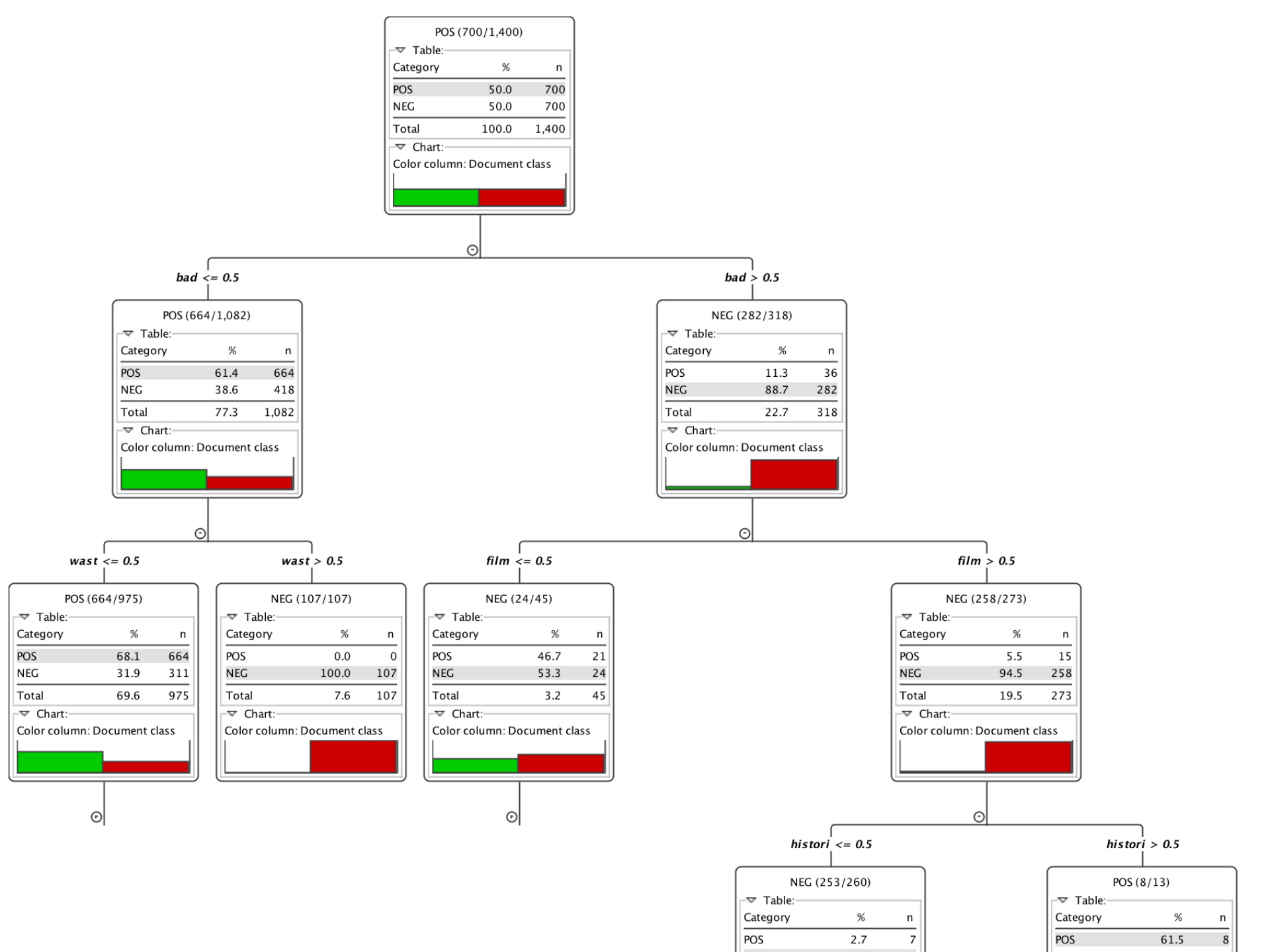

File Reader
Document Creation
Preprocessing
Document vector
Category to class
Color Manager
Read IMDb reviews from CSV file
Transformation of strings to documents
Preprocessing of documents
Create bit vectors for documents
Extract sentiment label
Color by sentiment label
Decision Tree Learner
ROC Curve
Partitioning
Build decision tree model
Decision Tree Predictor
Score decision tree model
Scorer
Training / test set
Apply decision tree model
Score decision tree model

File

Table "IMDb-sample.csv" - Rows: 2000 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	Index	URL	Text	Senti...
Row0	3617	http://www....	Girlfight follows a project dwelling New York high school girl from a sense of futility into the world of amat...	POS
Row1	3671	http://www....	Hollywood North is an euphemism from the movie industry as they went to Canada to make movies becau...	POS
Row2	3157	http://www....	That '70s Show is definitely the funniest show currently on TV. I started watching it about two and a half y...	POS
Row3	660	http://www....	9/10- 30 minutes of pure holiday terror. Okay, so it's not that scary. But it sure is fun.The Crypt Keeper (...	POS
Row4	265	http://www....	A series of random, seemingly insignificant thefts at her sister's boarding house has Miss Lemon quite agit...	POS
Row5	4027	http://www....	A very good adaptation of the novel by amrita pritam. Urmila and manoj bajpai have given their best.ther...	POS
Row6	5820	http://www....	Ah, Moonwalker, I'm a huge Michael Jackson fan, I grew up with his music, Thriller was actually the first m...	POS
Row7	1574	http://www....	Although the beginning of the movie in New York takes too long, the movie is a must see for people who li...	POS
Row8	10668	http://www....	As many reviewers here have noted, the film version differs quite a bit from the stage version of the story....	POS
Row9	1473	http://www....	Bear in mind, any film (let alone documentary) which asserts any kind of truth, will generate an adverse an...	POS
Row10	8337	http://www....	Being a big fan of the romantic comedy genre, and therefore having seen a large number of these films, it...	POS
Row11	11217	http://www....	Being an otaku since the days of Robotech, I can still say that Gunbuster is one of my favorite animes of all...	POS
Row12	12389	http://www....	Bored with the normal, run-of-the-mill staple films to watch this Halloween that I've seen over and over a...	POS
Row13	1212	http://www....	Care Bears Movie 2: A New Generation isn't at all a bad movie. In fact, I like it very much. Yes I admit the ...	POS
Row14	5272	http://www....	Deodato brings us some mildly shocking moments and a movie that doesn't take itself too seriously. Absol...	POS
Row15	9536	http://www....	Deranged and graphically gory Japanese film about little beings taking people over and turning them into ...	POS
Row16	11782	http://www....	Don't know if this contains any spoilers or not, but I don't want to risk being blacklisted until the year 346...	POS
Row17	11469	http://www....	Everyone knows about this "Zero Day" event. What I think this movie did that Elephant did not is that they ...	POS
Row18	9228	http://www....	Expecting to see another Nunsploitation movie with a mean Mother Superior abusing and torturing her cha...	POS
Row19	9945	http://www....	First things first! This isn't an action movie although there is a lot of action in it! I think you can compare it t...	POS
Row20	1721	http://www....	Had this movie been made a few years later, I would have given it a lower score. However, for 1909, this ...	POS
Row21	5992	http://www....	I admit I had no idea what to expect before viewing this highly stylized piece. It could have been the cure ...	POS
Row22	4678	http://www....	I chose to watch this film at Tribeca based on Judd Hirsch and Scott Cohen and found it to be one of the b...	POS
Row23	3422	http://www....	I desperately need this on a tape, not a DVD, and soon!I have one nephew who is in the infantry but has n...	POS
Row24	6823	http://www....	I have spent the last week watching John Cassavetes films - starting with 'a woman under the influence' an...	POS
Row25	6333	http://www....	I just got this video used and I was watching it last night. The acting started out extremely bad (hey-----...	POS
Row26	3976	http://www....	I just saw The Drugs Years on VH1 and I love it. I think it reflects the drug history very well and most impo...	POS
Row27	10202	http://www....	I personally liked this movie and am alarmed at the rating's some people have given it. It is a movie base...	POS
Row28	2844	http://www....	I saw 'New York: I Love You' today and loved it! I was really looking forward to seeing this after watching 'P...	POS
Row29	5169	http://www....	I saw the last five or ten minutes of this film back in 1998 or 1999 one night when I was channel-surfing ...	POS
Row30	10768	http://www....	I saw this ages ago when I was younger and could never remember the title, until one day I was scrolling t...	POS
Row31	6827	http://www....	I spent 5 hours drenched in this film. Nothing I have ever seen comes close to the delicious funk this film l...	POS

Row ID	Docu...	home	funniest	movi	funni	cast	joke	come	curious	continu	stori	consid	ve
1	""	1	1	1	1	1	1	1	1	1	1	1	1
2	""	0	0	1	0	1	0	0	0	0	0	0	0
3	""	0	0	1	0	0	0	0	0	0	0	0	0
4	""	0	0	0	0	0	0	0	0	0	0	0	0
5	""	0	0	1	1	1	1	0	0	0	0	0	0
6	""	0	1	1	0	0	0	0	0	0	0	0	1
7	""	0	0	1	0	0	0	0	0	0	0	0	0
8	""	0	0	0	0	0	0	0	0	0	0	0	0
9	""	0	0	1	1	0	0	0	0	0	0	0	0
10	""	0	0	1	0	0	0	0	0	0	0	0	1
11	""	0	0	0	0	0	0	0	0	0	1	0	0
12	""	0	0	0	0	0	0	0	0	0	1	0	1
13	""	0	0	1	0	0	0	0	0	0	0	0	0
14	""	0	0	1	0	0	0	0	0	1	0	0	0
15	""	0	0	1	0	0	0	0	0	1	1	0	0
16	""	0	0	1	1	0	0	1	0	0	1	0	0
17	""	1	0	0	0	0	0	0	0	0	0	0	0
18	""	0	0	1	0	1	0	0	0	0	1	0	0
19	""	1	0	1	0	0	0	0	0	0	0	0	0
20	""	1	1	1	0	0	0	0	0	0	0	0	0
21	""	0	0	1	0	0	0	0	0	0	0	0	0
22	""	0	0	1	0	0	0	0	0	0	0	0	0
23	""	0	0	0	1	0	1	0	0	0	0	0	0
24	""	0	0	0	0	0	0	0	0	0	0	0	0
25	""	0	0	0	0	0	0	0	0	0	0	0	0
26	""	0	0	0	0	0	0	0	0	0	1	0	0
27	""	0	0	1	1	0	0	0	0	0	1	1	0
28	""	0	0	1	1	0	0	1	0	0	0	0	0
29	""	0	0	1	0	1	0	0	0	0	0	0	0
30	""	0	0	1	0	0	0	0	0	0	1	0	0


POS (700/1,400)
Table:
Category % n
POS 50.0 700
NEG 50.0 700
Total 100.0 1,400
Chart:
Color column: Document class
bad <= 0.5
bad > 0.5
POS (664/1,082)
Table:
Category % n
POS 61.4 664
NEG 38.6 418
Total 77.3 1,082
Chart:
Color column: Document class
NEG (282/318)
Table:
Category % n
POS 11.3 36
NEG 88.7 282
Total 22.7 318
Chart:
Color column: Document class
wast <= 0.5
wast > 0.5
film <= 0.5
film > 0.5
POS (664/975)
Table:
Category % n
POS 68.1 664
NEG 31.9 311
Total 69.6 975
Chart:
Color column: Document class
NEG (107/107)
Table:
Category % n
POS 0.0 0
NEG 100.0 107
Total 7.6 107
Chart:
Color column: Document class
NEG (24/45)
Table:
Category % n
POS 46.7 21
NEG 53.3 24
Total 3.2 45
Chart:
Color column: Document class
NEG (258/273)
Table:
Category % n
POS 5.5 15
NEG 94.5 258
Total 19.5 273
Chart:
Color column: Document class
histori <= 0.5
histori > 0.5
NEG (253/260)
Table:
Category % n
POS 2.7 7
POS (8/13)
Table:
Category % n
POS 61.5 8

Document ...	NEG	POS
NEG	291	9
POS	29	271

Correct classified: 562

Wrong classified: 38

Accuracy: 93.667 %

Error: 6.333 %

Cohen's kappa (κ) 0.873

Ďakujem za pozornosť!

257890@mail.muni.cz