Ekstrakcja emocji z tekstu

Przetwarzanie języka naturalnego
w kontekście wyborów – i nie tylko.

Aleksander Wawer

stonecenter

Agenda

- Stone Center i General Inquirer
- Przetwarzanie języka a wybory
- Sentiment analysis jako klasyfikacja i "bag of words"
- "Bag of words" a machine learning
- Systemy oparte o reguły i parsowanie. Korpus emocji w tekście.

General Inquirer

Korzenie *content analysis*

stonecenter

Philip Stone

- Philip J. Stone – (zm. w styczniu 2006) był profesorem psychologii pozytywnej na Harvard University w USA i pionierem komputerowej *content analysis*.
- Autor i początkowo główny developer (TrueBasic na mainframe IBMa) pionierskiej aplikacji General Inquirer.

Stone Center

- Lokalizacja: Chorwacja, Mali Losinj. (http://info.stone-center.eu)
- Cel: kontynuacja zainteresowań badawczych i niektórych projektów rozpoczętych przez prof. Stone.
- Laboratorium Content Analysis
  - Udostępnianie General Inquirer dla wspólnoty akademickiej.
  - Dalszy rozwój narzędzia i słowników.
  - Zastosowanie w kontekstach badawczych.
  - Rozwój GI i *sentiment analysis* w wielu językach.

stonecenter

General Inquirer

- Content analysis – analiza treści przekazu
- "Who says what, to whom, why, to what extent and with what effect?" [Laswell].
- Systematyczna, obiektywna, ilościowa analiza własności przekazu [Kiberly Neuendorf].
- Wnioskowanie poprzez systematyczną i obiektywną identyfikację określonych własności tekstu [Ole Holsti, Stone].

General Inquirer

- Słowniki. Macierz 182 kategorie * 11767 znaczeń, m.in.:
  - Semantyczny dyferencjał i 6 polarnych kategorii C. Osgooda pozytywne/negatywne; aktywne/pasywne; silne/słabe.
  - 7 kategorii słów przyjemności, bólu, zalet i wad (charakteru).
  - 9 kategorii T. Parsonsa (ekonomia, prawo, polityka, religia itd.).
  - 17 kategorii orientacji kognitywnej (nastawienie na rozwiązywanie problemów, porównanie, ocena itd.).
  - Słownik wartości politycznych H. Laswella.
  - I wiele innych.
- Stemmer. Oryginalnie słownikowy, obecnie hybrydowy: częściowo algorytmiczny.
- Dezambiguator oparty o reguły.

Dezambiguator GI - składnia

* subrule	code	usage#	comment
* WOR	5W-O	1201	Word *or* test
* WORK	5WKO	79	Word *or* test skipping key word
* WAND	5W-A	111	Word *and* test, with words in order
* WANDK	5WKA	25	All test words in order, skipping key word
* WADJ	5W-N	5	Words must be in order with no intervening words
* TOR	5T-O	3958	Tag *or* test
* TORK	5TKO	124	Tag *or* test skipping key word
* TAND	5T-A	13	Tag *and* test
* TANDK	5TKA	7	Tag *and* test skipping key word
* TSAMEK	5TKS	0	All tags must be on the same word, key word not tested
* TSAME	5T-S	12	All tags must be on the same word
* TADJ	5T-N	16	No intervening words between tags
* TSAMEM	5T-SM	164	First tag must be on a word without the other tags
* TSAMEMK	5TKSM	2	First tag must be on a word without the other tags
* SUPV	2S	475	Transfers to *SUPV* routine
* GOTO	2G	143	Transfer to rules for another word

Przykład reguły - ADVANCE

```
ADVANCE          #25
  5  TOR(K+0,K+0,APLY(1),,ING.
  6  TOR(K+0,K+0,,10,ED.
  7  TOR(K-1,K-1,APLY(3),,DET.PREP.LY.EVAL.
  8  TSAMEM(K-1,K-1,APLY(3),APLY(1),PUNC.COMMA.DASH.
 10  WOR(K-1,K-1,DELID(4),,IN.
 11  TOR(K-1,K-1,APLY(1),,TO.MOD.PRON.DO.NEG.
 12  TOR(K+1,K+1,APLY(1),,DET.
 13  TSAMEM(K+1,K+1,APLY(1),APLY(2),PRON.DEF1.
```

GI jako „gold standard”

The labeled dataset used as a gold standard is General Inquirer lexicon (Stone et al., 1966) as in the work by Turney and Littman (2003). We extracted the words tagged with “Positiv” or “Negativ”, and reduced multiple-entry words to single entries.

Takamura H., Inui T., Manabu O.. *Extracting Semantic Orientations of Words using Spin Model*. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics 2005

stonecenter

Przetwarzanie języka naturalnego a wybory

Kontekst polityczny

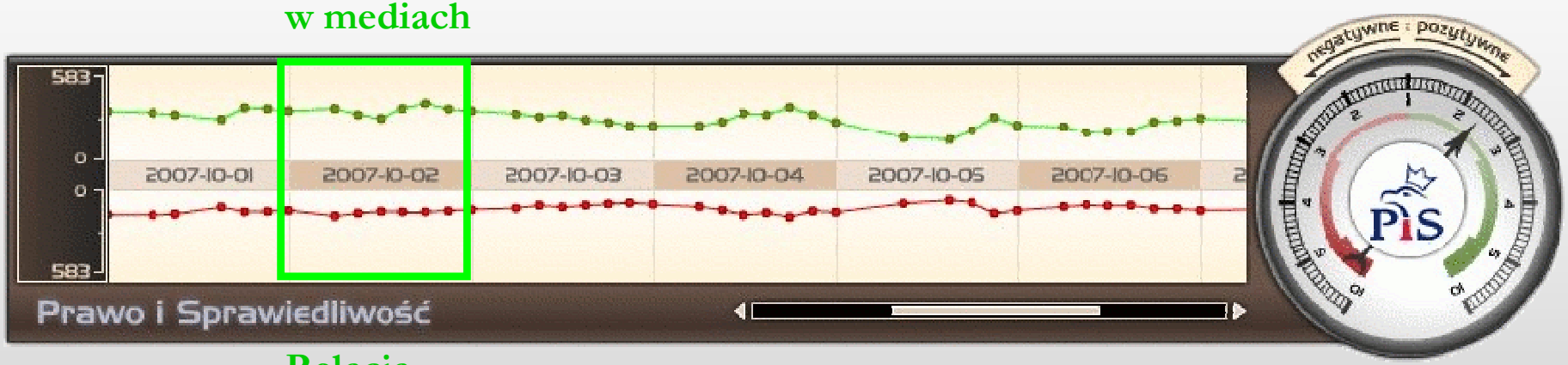
stonecenter

Polska – Barometr Polityczny

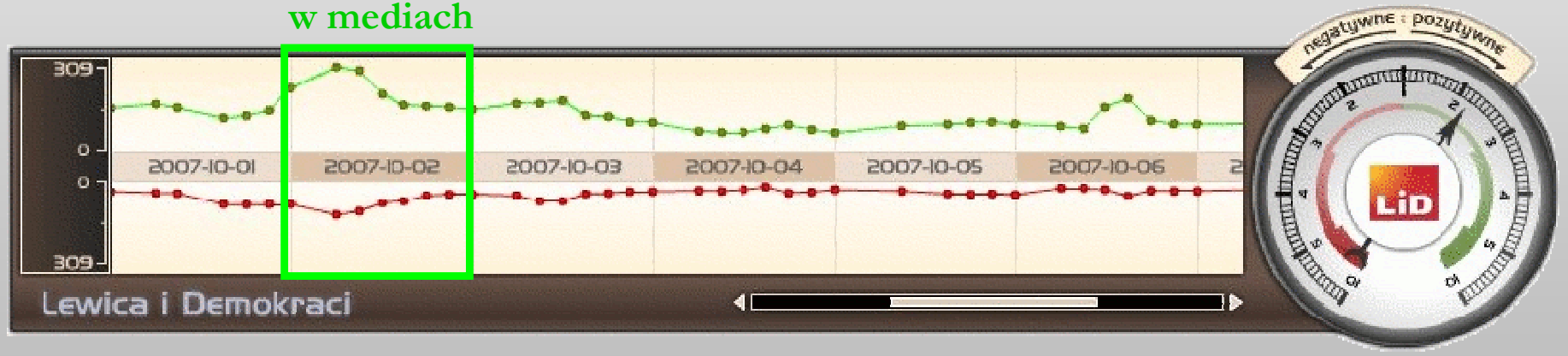
- Prosty *sentiment analyzer* (http://barometr-polityczny.pl) oparty o ontologię partii politycznych i należących do nich polityków.
- Co 3 godziny automat pobiera z Internetu 300-400 artykułów dotyczących bieżącej sytuacji politycznej w kraju. Poddawane są one analizie pod kątem występowania terminów nacechowanych pozytywnie lub negatywnie, występujących w sąsiedztwie nazw partii i nazwisk kluczowych polityków.
- Leksykon stworzony przy wykorzystaniu Korpusu IPI PAN oraz słowników General Inquirer.

Kto wygrał debatę Kwaśniewski-Kaczyński?

Relacja
w mediach



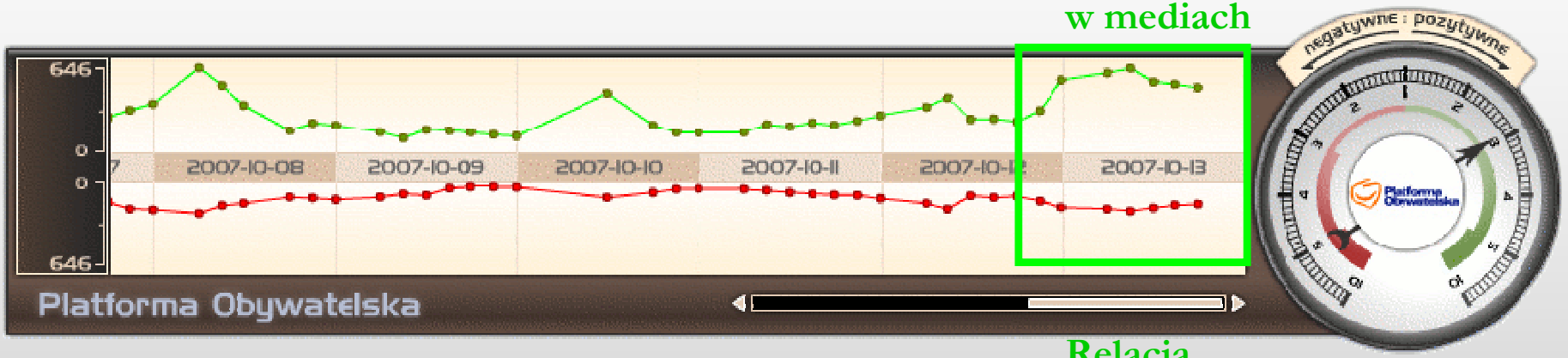
Relacja
w mediach



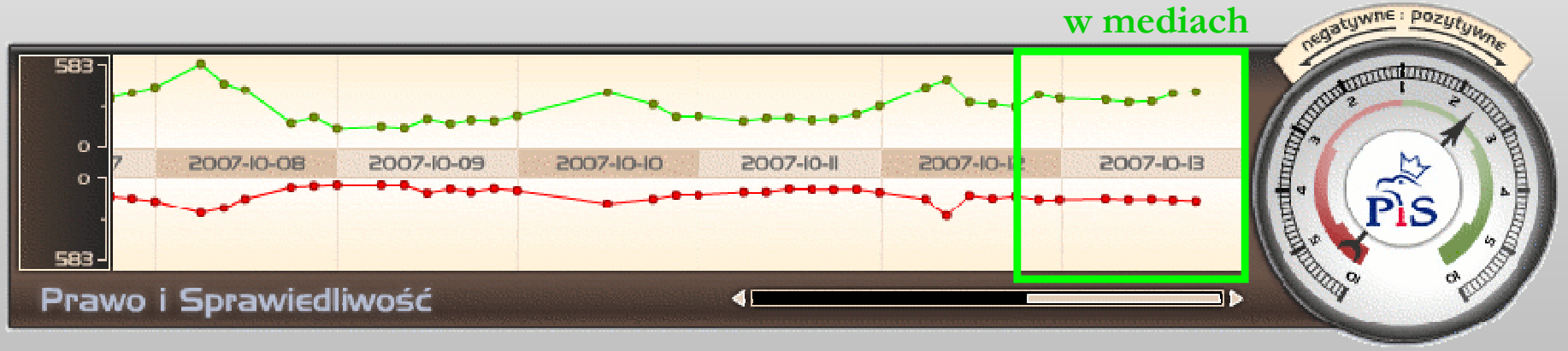
stonecenter

Kto wygrał debatę Kaczyński-Tusk?

Relacja w mediach



Relacja w mediach



stonecenter

Czy można było przewidzieć G.W. Busha w 1987 roku?

- Dane: tekst z 33 platform (*formal written statement of party's principles and goals*) partii Republikańskiej i Demokratycznej z lat 1844 do 1964.
- Wektory z odsetkami, jakie stanowiły słowa danej kategorii we wszystkich tekstach danej platformy.
- Około 80% wśród 42 kategorii słownika Laswella zmieniało się w sposób **cykliczny**.

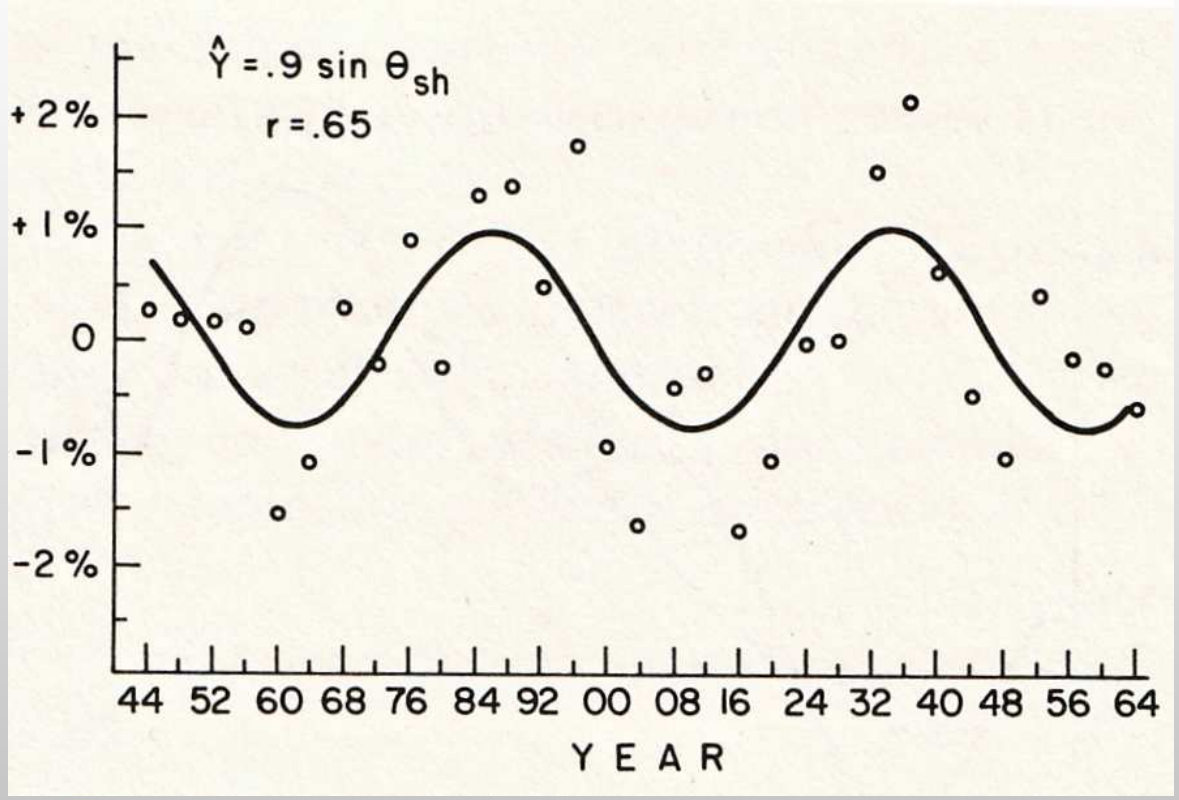
Cykliczność długo- i krótkoterminowa

- Krótki cykl – około 48 lat. Dynamika związana z ekonomią i doraźną orientacją polityczną.
- Długi cykl– około 150 lat. Systemy społeczne.

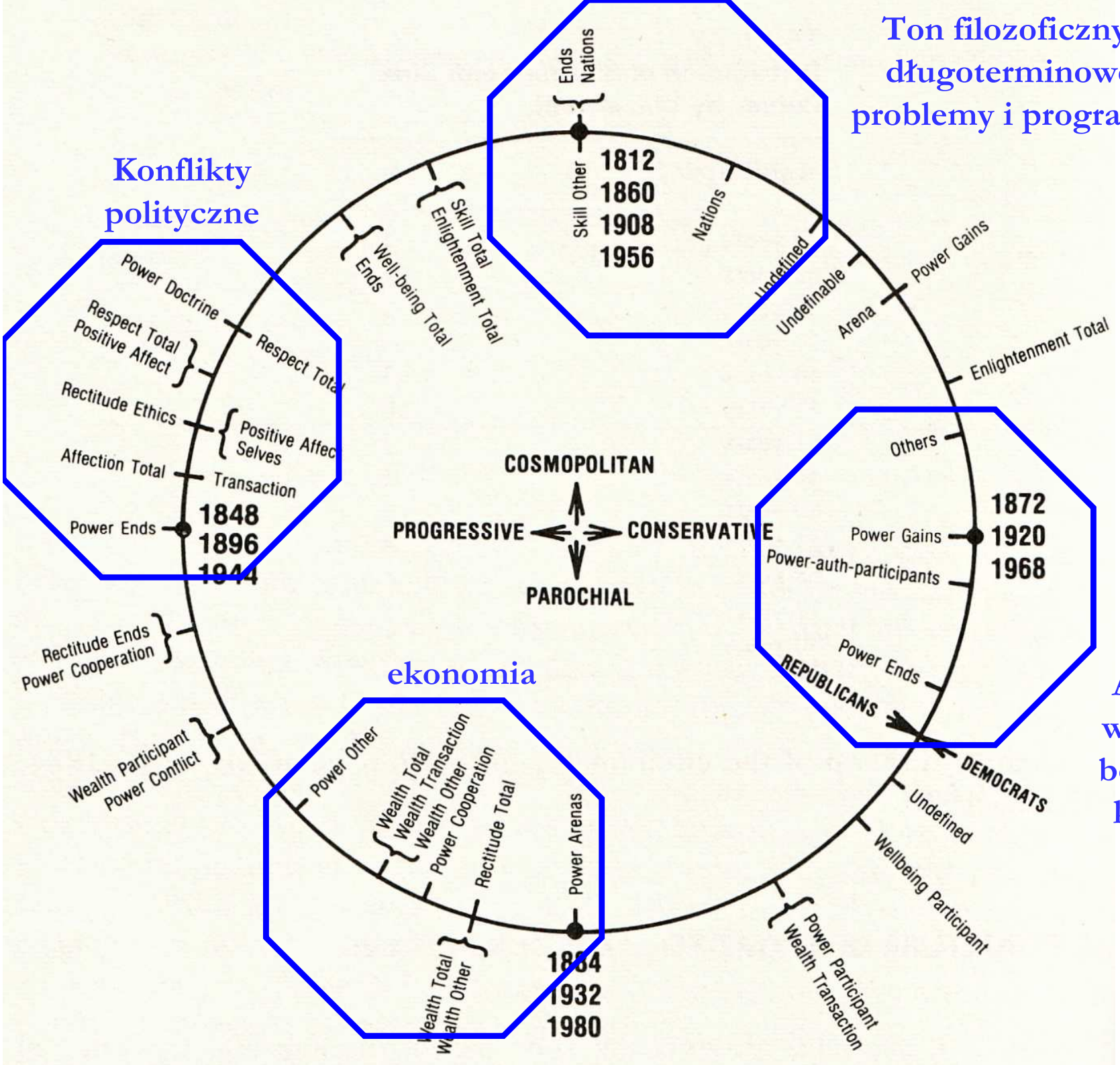
Liczba kategorii	**Demokraci**		
Republikanie		Z cykliczną zmianą	Bez cyklicznej zmiany
	Z cykliczną zmianą	12	12
	Bez cyklicznej zmiany	8	10

stonecenter

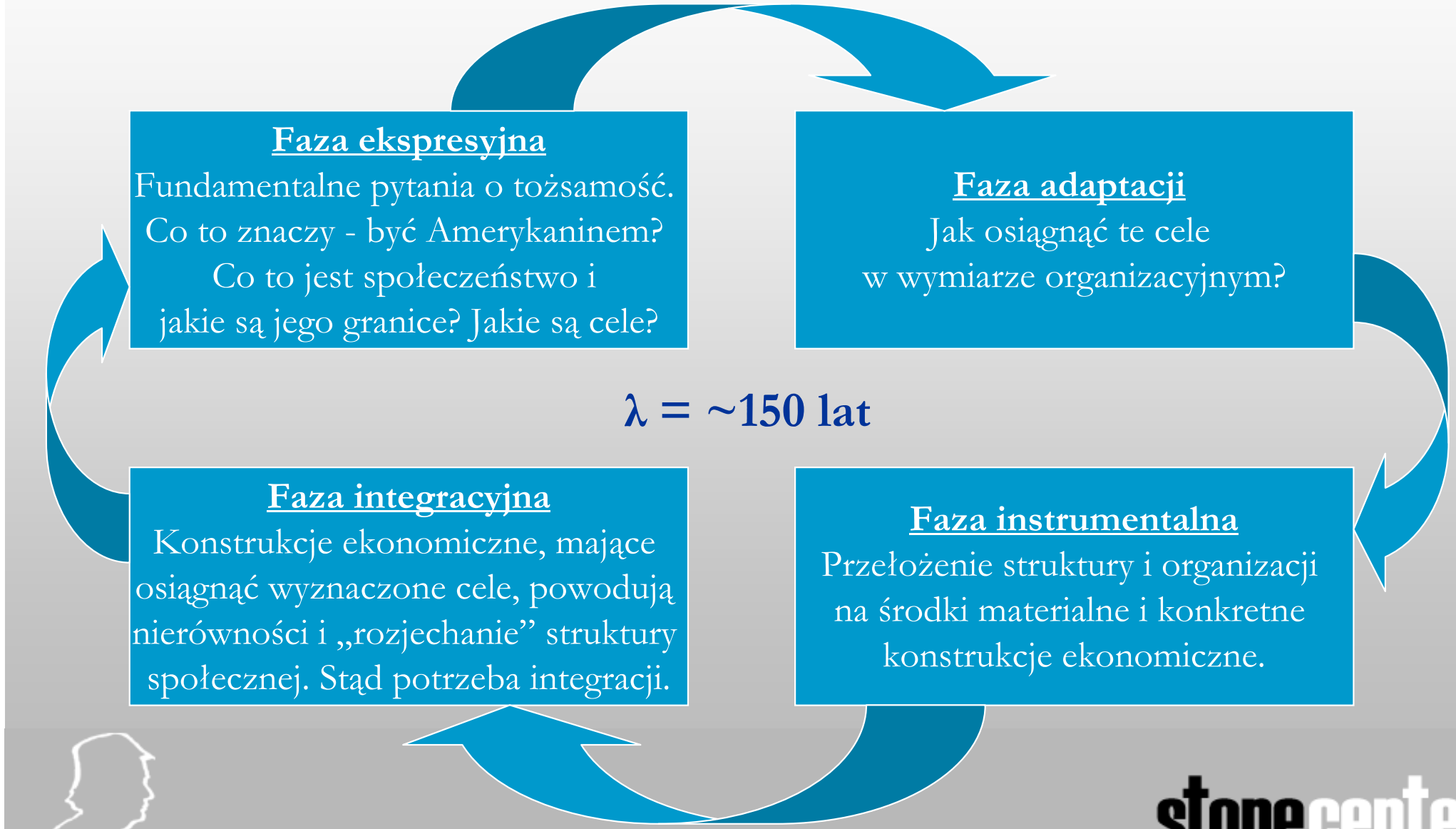
Przykład krótkoterminowego cyklu



Krótkoterminowy cykl – dopasowanie kategorii WEALTH w platformach partii demokratycznej


Ton filozoficzny, długoterminowe problemy i programy
Konflikty polityczne
ekonomia
Autorytatywna władza, agencje bezpieczeństwa, konserwatyzm
COSMOPOLITAN
PROGRESSIVE
CONSERVATIVE
PAROCHIAL
Ends
Nations
Skill Other
1812
1860
1908
1956
Nations
Undefined
Undefinable
Power Gains
Arena
Enlightenment Total
Skill Total
Enlightenment Total
Well-being Total
Ends
Power Doctrine
Respect Total
Positive Affect
Respect Total
Rectitude Ethics
Positive Affect
Selves
Affection Total
Transaction
Power Ends
1848
1896
1944
Others
1872
1920
1968
Power Gains
Power-auth-participants
Power Ends
REPUBLICANS
DEMOCRATS
Undefined
Wellbeing Participant
Power Participant
Wealth Transaction
Rectitude Ends
Power Cooperation
Wealth Participant
Power Conflict
Power Other
Wealth Total
Wealth Transaction
Wealth Other
Power Cooperation
Rectitude Total
Power Arenas
Wealth Total
Wealth Other
1884
1932
1980

Cykliczna dynamika kultury



stonecenter

Sentiment analysis jako klasyfikacja

Podejście „bag of words”.

stonecenter

Sentiment analysis jako klasyfikacja

- Analizujemy nastawienie dokumentu jako **całości**, np. czy opinia o produkcie (*product review* jako cały dokument) jest pozytywna, czy negatywna? Nie szukamy przedmiotu opinii ani jej posiadacza.
- Metody machine learning (Naive Bayes, maksymalna entropia, SVM) działają w tym zastosowaniu **gorzej** niż dla analogicznego problemu automatycznej kategoryzacji dokumentów ze względu na temat [Pang et al. 2002].
- Powód: klasyfikacja tekstów ze względu na temat jest bardziej podatna na słowa kluczowe.

Problem z „bag of words” (1)

- Przymiotnik „*unpredictable*” może mieć znaczenie negatywne w „*unpredictable steering*” (review samochodu) i pozytywne jako „*unpredictable plot*” (review filmu).
- “*How could anyone sit through this movie?*” – nie ma ani jednego negatywnego słowa. Rozumienie wymaga poznania kontekstu. FrameNet? Ontologie (CYC?)
- „*There isn't anything that I disliked about the product*” – 3 negatywne słowa! Reguły, parsowanie?

Problem z „bag of words” (2)

- Czy istnieje zbiór słów, których użycie jest wystarczające do oddzielenia opinii pozytywnych od negatywnych?
- Dane: 700 pozytywnych i 700 negatywnych *reviews*.

	Zbiory słów	Dokładność
Osoba 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58%
Osoba 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64%

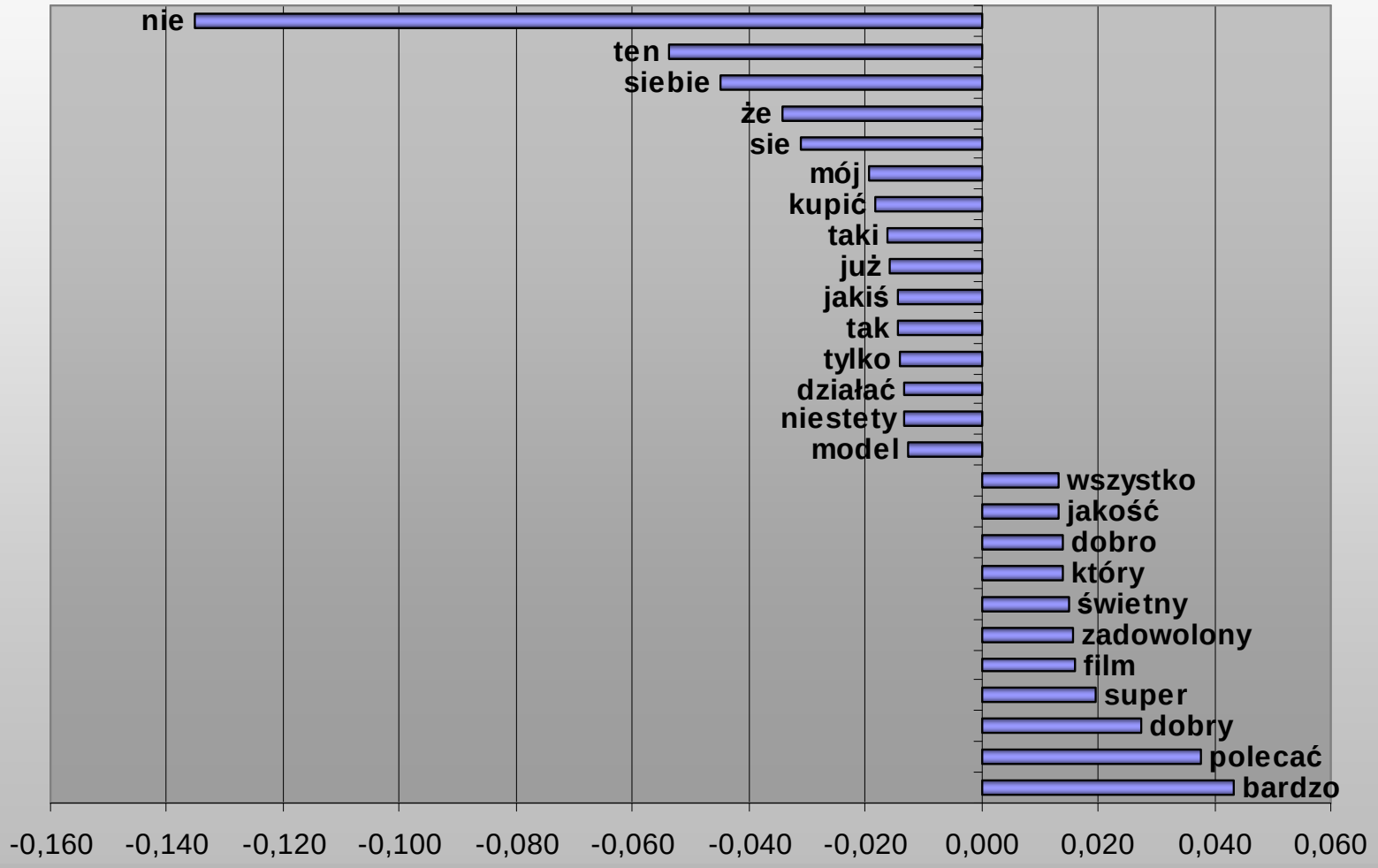
Źródło: Pang et al. 2002

stonecenter

Leksykony z *product reviews* - teoria

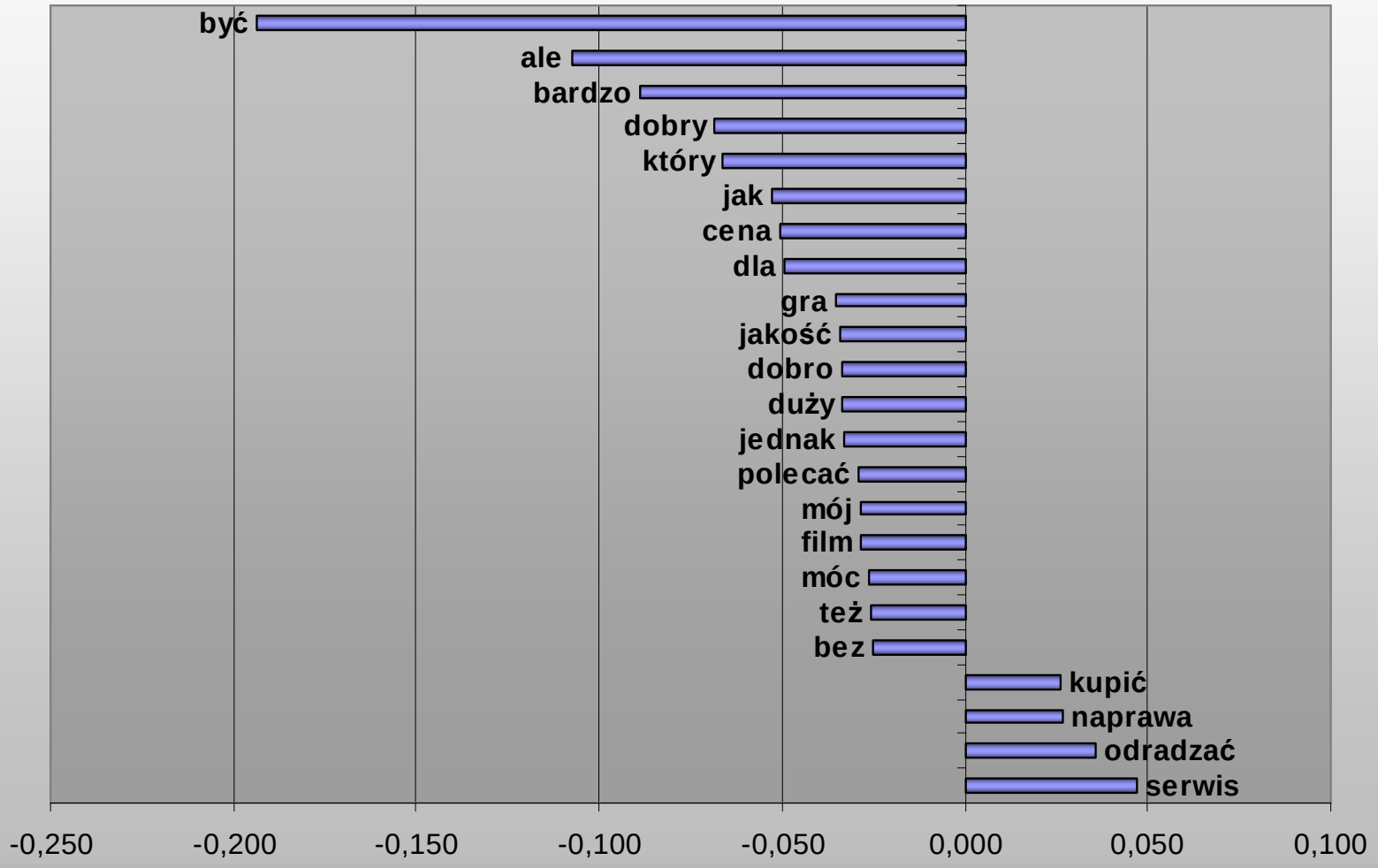
- Czy możliwe jest automatyczne wywnioskowanie emocjonalnie nacechowanych słów z polskich *product reviews*?
- Korpus: ponad 11 tys. opinii (*reviews*) z polskiego Internetu. Każda z opinii posiada przypisaną walencję, od 0 do 5 gwiazdek.
- Prawo Zipf'a. Częstość występowania n-tego najczęstszego słowa (ranga) =~ 0.1/n (dla n<1000).
- Intuicja: w pozytywnych *product reviews* słowa pozytywne będą pojawiały się częściej. I vice versa.
- Porównanie częstości (rang) występowania słów z 3 korpusów:
  - Całego zbioru 11 tysięcy opinii (punkt odniesienia)
  - Podzbioru opinii negatywnych (0 i 1 gwiazdka)
  - Podzbioru opinii pozytywnych (4 i 5 gwiazdek)

Leksykony z *product reviews* – praktyka (pozytywne vs całość)



stonecenter

Leksykony z *product reviews* – praktyka (negatywne vs całość)



stonecenter

Wniosek – problem z „bag of words” (3)

- Zasób leksykalny wykorzystywany do recenzji produktów (*product reviews*) jest ubogi. Emocje w polityce wyrażają zupełnie inne terminy i wyrażenia.
- Wskazane jest wykorzystanie bardziej złożonych narzędzi. Absolutnym minimum jest detekcja inwersji, czyli: „polecam” vs „nie polecam”.

GI+C99+SVM

Sentiment analyzer przy użyciu
General Inquirera i machine learning

stonecenter

Sentiment analyzer

- Stworzyć zbiory tekstów treningowych z kilku domen. Oznaczyć (otagować) je ręcznie jako pozytywne/negatywne/neutralne - zmienna zależna.
- Wykonać wstępne przetwarzanie tekstów. Może to być:
  - Segmentacja (ekstrakcja fragmentu tekstu, który związany jest z danym przedmiotem / opisuje go). Tu C99 może okazać się przydatny.
  - Ustalanie referencji zaimków (anafora) w sąsiednich zdaniach. Dalsze poprawienie segmentacji.
  - Feature selection, skalowanie lub inne przekształcenia na wektorach GI przed wykorzystaniem ich w charakterze treningowego zbioru danych dla SVM.
  - Parsowanie powierzchniowe, identyfikacja zwrotów i fraz odnoszących się do przedmiotu na poziomie syntaktycznym.
  - N-gramy tagów GI jako alternatywny input GI.

GI+C99+SVM

- Stworzyć system odtwarzający ludzką klasyfikację tekstów w sposób niezależny od domeny.
- Teksty podzielone na dwie klasy:
  - pozytywne/zachęcające
  - negatywne/zniechęcające
- Dwie domeny:
  - *product reviews* o skrajnej polaryzacji, tzn. 0-1 lub 4-5 gwiazdek
  - posty z forum dyskusyjnego zachęcające lub zniechęcające do zażywania narkotycznych środków przeciwbólowych, opartych o opiaty.
- Zbiór 300 tekstów, po połowie negatywnych i pozytywnych.
- Jest to prosty eksperyment. Niewielka ilość tekstów, brak tekstów neutralnych (2 klasy zamiast 3).

Analiza wariancji, PCA

- Czy możemy przypisać decyzję podjętą przez człowieka do pewnych własności 180-wymiarowego zbioru danych, wektorów z GI?
  - Analiza wariancji wskazuje na brak istotnych statystycznie zależności między zmienną zależną (pozytywny lub zachęcający vs. negatywny lub zniechęcający) a 180 wymiarami z GI, zmiennymi niezależnymi.
  - Analiza głównych składowych (PCA) ujawnia 4 składowe stosunkowo łatwe w interpretacji oraz 1 „zaszumioną" składową, która niestety pokrywa 80% danych.

Przetwarzanie wstępne - C99

- Posty z forum dyskusyjnego poruszały niekiedy kilka tematów. Często tylko fragment tekstu zachęcał lub zniechęcał.
- Pytanie: jak zidentyfikować relewantną część wypowiedzi? Możliwa odpowiedź: segmentacja tekstu.
- Intuicja: segmenty tekstu z podobnym słownictwem prawdopodobnie dotyczą tego samego tematu.

C99 [Choi 2000]

- Niech $f_{i,j}$ oznacza częstotliwość występowania **kategorii** GI (oryg: słowa) *j* w zdaniu *i*. Dla każdej pary zdań generujemy macierz podobieństwa:

$$sim(x,y) = \frac{\sum_j f_{x,j} \times f_{y,j}}{\sqrt{\sum_j f^2_{x,j} \times \sum_j f^2_{y,j}}}$$

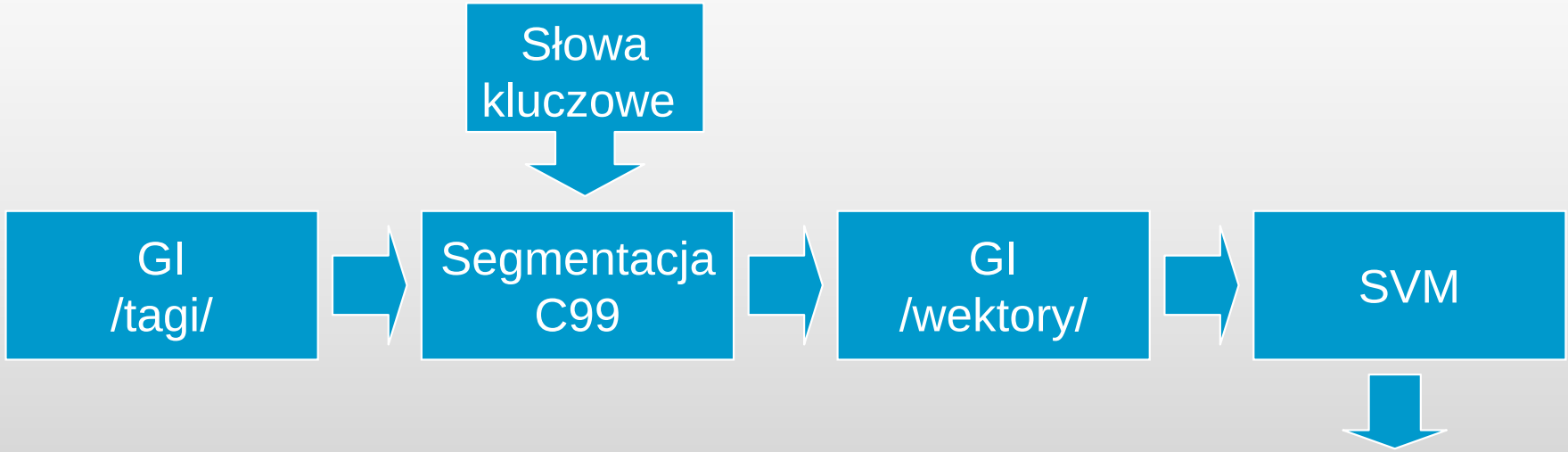
- Tworzymy macierz rang R_{mxm} – jako liczbę sąsiadów z niższym podobieństwem (sim), dla każdej pary zdań.

- Clustering top-down. Kryterium podziału danego segmentu B na podsegmenty $b_1, .. , b_m$ oparte jest o maksymalizację gęstości D, wyliczanej dla każdego potencjalnego podziału; s_k i a_k to odpowiednio suma rang i obszar k-tego segmentu.

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k}$$

stonecenter

GI+C99+SVM



Decyzja czy tekst jako całość
jest zachęcający czy zniechęcający
w odniesieniu do słowa kluczowego

stonecenter

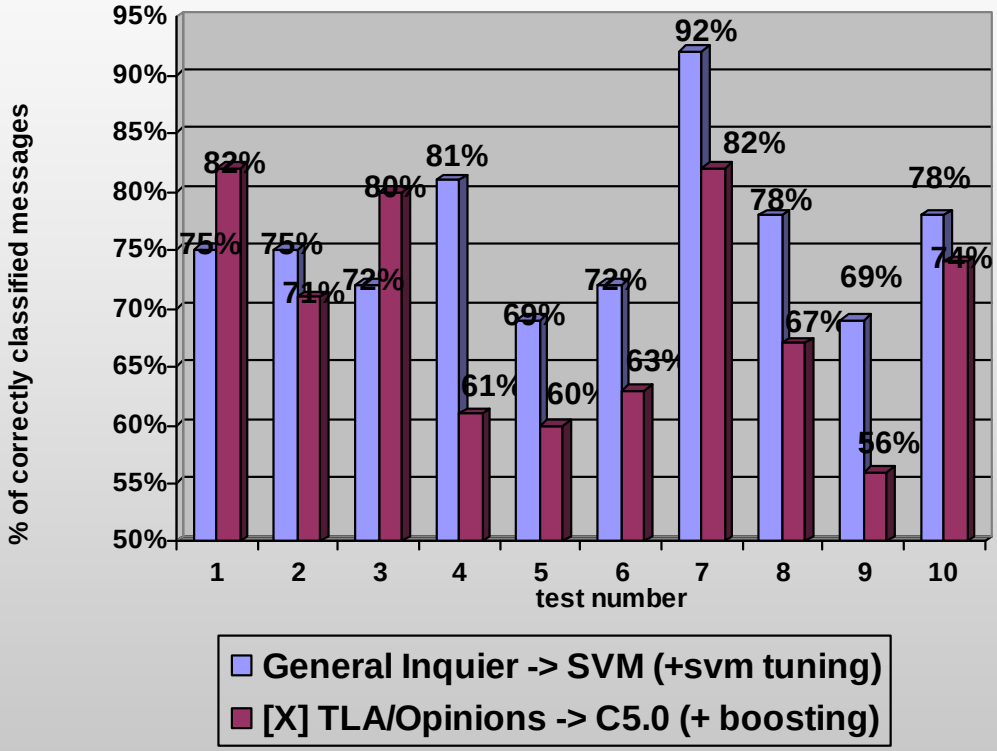
GI+C99+SVM vs [X]+C5.0

Dane dzielone losowo na 2 zbiory, treningowy (85%) i testowy (15%).

Liczymy dokładność jako odsetek dokumentów zaklasyfikowany przez SVM do prawidłowej klasy. Eksperyment powtarzamy 10 razy.

Wyniki:

- [X] oraz C5.0 osiągnął średnią dokładność **69.5%**; najgorszy przypadek 55,6% - prawie losowy.
- GI oraz SVM osiągnął średnią dokładność **76.1%**; najgorszy przypadek 69.4%.



stonecenter

Komentarz

- Czy SVM jest lepszą metodą niż C5.0? Czy GI bogatszym zasobem leksykalnym niż [X]?
- [Pang et al 2002] osiągnął dokładność SVM powyżej 80% dla modelu opartego o obecność unigramów; SVM trenowany na rec.arts.movies.reviews
- Czy zbiór danych Panga (*reviews*) był bardziej spójny i czysty niż posty o przeciwbólowych opiatach?
  A może obecność unigramów jest lepszym predyktorem niż wektory frekwencji kategorii GI?

Bibliografia [1]

- Choi F., *Advances in domain independent linear text segmentation*. Proceedings of NAACL-00 (2000)
- Pang B., Lee L., Vaithyanathan S., *Thumbs up? Sentiment Classification using Machine Learning Techniques*. EMNLP (2002)
- Stone P., Dunphy D., Smith M., Ogilvie D., *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press (1966)

Reguły i parsowanie

stonecenter

Systemy oparte o reguły – SA [Yi et al. 2003]

- (1) Ekstrakcja terminów opisujących cechy produktów (3 heurystyki wyboru potencjalnych terminów, potem Mixture Model [Zhai, Laffertry]).
  - **Digital Camera:** camera, picture, flash, lens, picture quality, battery (..)
- (2) Leksykon (GI+WordNet+DAL, ok. 2500 przymiotników i 500 rzeczowników)
- (3) Baza reguł w liczbie około 120.

Systemy oparte o reguły – SA [Yi et al. 2003]

I am impressed by the flash capabilities.
- pattern : "impress" + PP(by;with)
- subject : flash

SA identifies the *T-expression* of the sentence: <flash capability, impress, ""> and infers that the *target* (PP lead by “by” or “with”), the flash capabilities, has positive sentiment:
(flash capability, +).

This camera takes excellent pictures.
- matching sentiment pattern: <"take" OP SP>
- subject phrase (SP): this camera
- object phrase (OP) : excellent pictures
- sentiment of the OP: positive

SA identifies the *T-expression* of the sentence: <camera, take, excellent picture> and infers that the sentiment of *source* (OP) is positive, and associates positive sentiment to the *target* (SP): (camera, +).

stonecenter

Korpus emocji

- Do czego potrzebny jest anotowany korpus emocji w tekście [Wiebe et al. 2005]?
  - Jako treningowy i testowy zbiór danych dla *machine learning* (benchmark różnych algorytmów i implementacji).
  - Automatyczne stworzenie (wywnioskowanie) zbioru reguł.
  - Wsparcie dla systemów *multi-perspective question answering*.

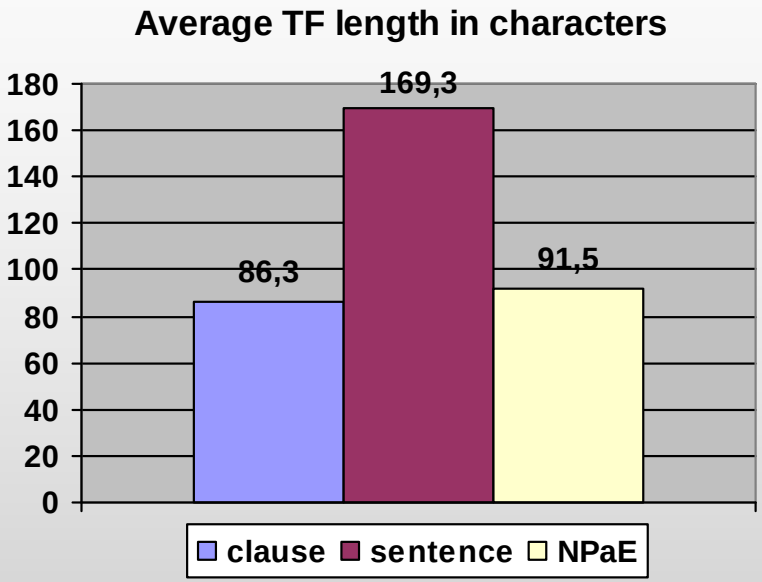
Korpus MPQA

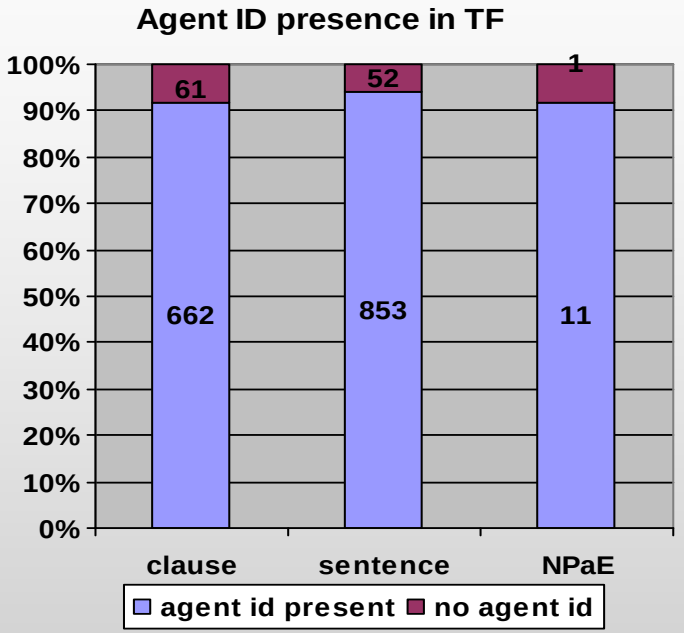
- Korpus MPQA (Multi-Perspective Question Answering Opinion Corpus) jest zbiorem anglojęzycznych newsów z prasy światowej.
- Teksty pochodzą ze 187 źródeł z kilku krajów. Datowane są pomiędzy czerwcem 2001 a majem 2002.
- Prace związane ze zbieraniem tekstów oraz anotacją wykonano w ramach Workshop on Multi-Perspective Question Answering (2002) sponsorowanego przez ARDA.

Anotacje MPQA

- MPQA zawiera cztery rodzaje anotacji [Wiebe et al., 2005]: *direct-subjective*, *expressive-subjective*, *objective-speech-event* oraz *agent annotation*.
- W ramach tych anotacji istnieją dwa atrybuty, wyrażające referencję do jakiegoś obiektu (*entity*).
- Dla anotacji *direct-subjective* (*direct mentions of private states and speech events expressing private states*) atrybut "attitude-toward" – agent, na którego skierowana jest opinia.
- Dla anotacji *agent annotation* (*phrases that refer to sources of private states and speech events, or phrases that refer to agents who are targets of an attitude*) atrybut "nested-target" – wykorzystywany gdy anotowany agent jest celem pozytywnej lub negatywnej opinii.
- „Agent" jest albo źródłem komunikatu (wypowiadającym zdanie) albo też tym, który doświadcza prywatny stan. W korpusie MPQA „agenci" to zazwyczaj osoby lub kraje. W naszej analizie ograniczyliśmy się do nazw krajów.

Wyniki – komercyjny entity extractor vs. MPQA (1)





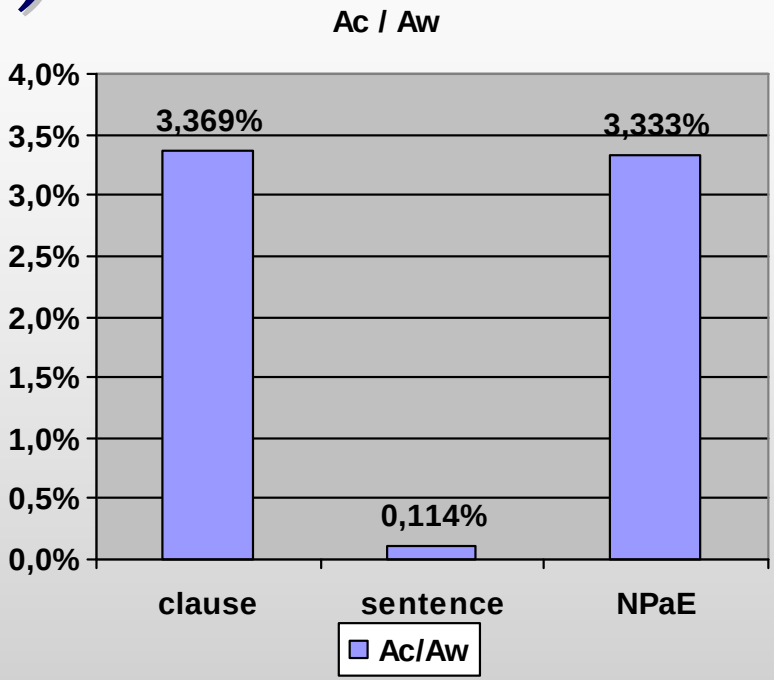
Agent ID powinien znajdować się w obrębie tekstu, jaki zwróciła reguła

TF = output reguły

stonecenter

Wyniki – komercyjny entity extractor vs. MPQA (2)

- Chcemy minimalizacji Ac/Aw.
- Najlepiej działa reguła „sentence”.



Aw – słowa zawierające ekspresję opinii, zawarte w całości w tekście, zwróconym przez regułę.

Ac – słowa zawierające ekspresję opinii, które wykroczyły częściowo lub w całości poza tekst, zwrócony przez regułę.

stonecenter

Komentarz

- Ze względu na przedstawione (proste) miary, najlepszą regułą wydaje się „sentence”. Niestety jest to reguła dopuszczająca najwięcej szumu – słów i zwrotów nie związanych z wyrażeniami opinii.
- Nawet reguła NPaE (Noun Phrase around Entity) nie jest do końca precyzyjna. A takich reguł powinniśmy stworzyć kilkadziesiąt.
- Na ile precyzyjne są reguły w SA oraz [X]? Ile dopuszczają *false positives*?
- Swoją drogą, uzyskanie wyższej precyzji wymaga ustalenia referencji zaimków osobowych.

Bibliografia [2]

- Wiebe J., Wilson T., Cardie C., *Annotating Expressions of Opinions and Emotions in Language.* Language Resources and Evaluation. 39(2-3):165–210. (2005)
- Yi J., Nasukawa T., Bunescu R., Niblack W., *Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques.* Proceedings of the 3rd IEEE International Conference on Data Mining (2003)

Wyzwania

- Żaden komercyjny ani niekomercyjny system –znany autorowi tej prezentacji – nie wykorzystuje głębokiego parsowania. HPSG a sentiment analysis?
- Wyjść poza jednowymiarową analizę (pozytywne-negatywne). Osgood i nie tylko.
- Analiza innych domen niż tylko *product reviews*. Emocje w kontekście politycznym, życiu publicznym, na blogach, w biznesie, sporcie itd..
- Zbudować korpus emocji. Nadbudowa nad korpusem IPI?