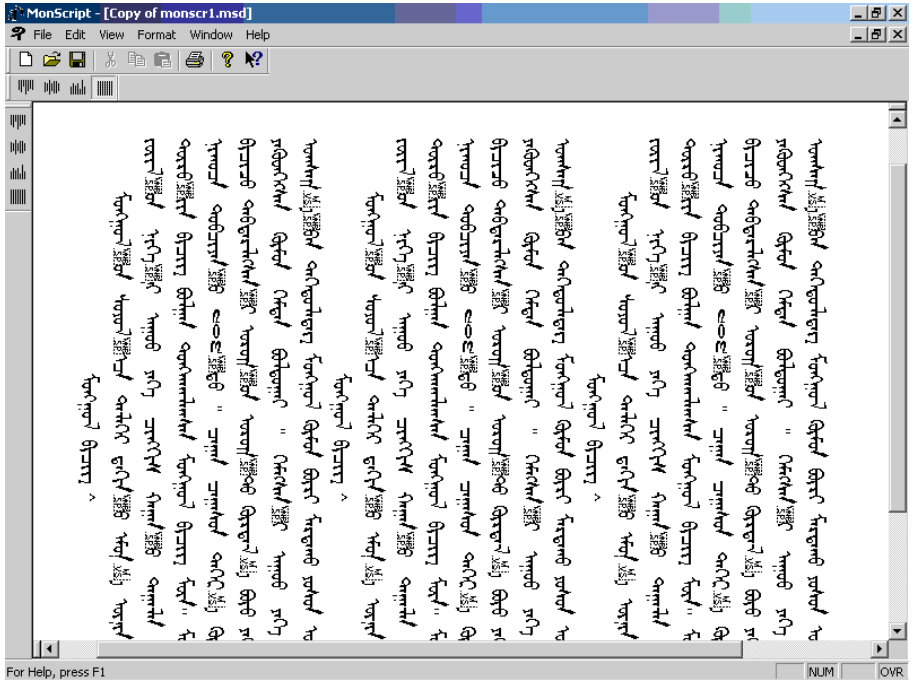


Хандлагын шинжилгээ

Хандлагын үгийн сан

Ерөнхий лавлах

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Веб хуудас: http://www.wjh.harvard.edu/~inquirer
- Категорийн жагсаалт: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Хүснэгт: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls
- Катеориуд:
  - Эерэг (1915 үгс) ба Сөрөг (2291 үгс)
  - Хүчтэй ба сул, Идэвхтэй ба Идэвхгүй, Хэт их ба Доогуур
  - Таашаал, өвдөлт, зан чанар, орлогч, сэдэл, танин мэдэхүйн чиг баримжаа гэх мэт.
- Судалгаанд ашглахад үнэгүй

LIWC (Хэл шинжлэлийн лавлах ба үгийн тоо)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

- Веб хуудас: http://www.liwc.net/
- 2300 үгс, >70 анги
- **Сэтгэл хөдлөлийн процесс**
  - сөрөг эмоци (*муу, хачин, үзэн ядалт, асуудал, хүнд хэцүү*)
  - эерэг эмэци (*сайхан, аятайхан*)
- **Танин мэдэхүйн процесс**
  - Таамаглах (магадгүй*, тааварлах*), Хориглолт (*хориглох, хязгаарлах*)
- **Төлөөний үг, Үгүйсгэл** (үгүй*, хэзээ ч үгүй*), **Тоон үзүүлэлтүүд** (цөөн*, олон*)
- $30 эсвэл $90 төлбөр

MPQA Субъектив хэлц үгийн сан

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

- Веб хуудас: http://www.cs.pitt.edu/mpqa/subj_lexicon.html
- 8221 леммагаас 6885 үгс
  - 2718 эерэг
  - 4912 сөрөг
- Үг бүрт (хүчтэй, сул) эрчимжилт тэмдэглэсэн
- GNU GPL

Bing Liu үзэл бодлын үгийн сан

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.

- Bing Liu-ийн веб хуудсын үзэл бодол олборлолт
- http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

- 6786 үгс
  - 2006 эерэг
  - 4783 сөрөг

SentiWordNet

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

- Веб хуудас: http://sentiwordnet.isti.cnr.it/
- WordNet –ийн бүх synsets –ийг эерэг, сөрөгийн зэрэг, ба төвийг сахисан/бодит байдлыг автоматаар тэмдэглэдэг
- [estimable(J,3)] “may be computed or estimated”

  Эерэг 0 Сөрөг 0 Бодит 1
- [estimable(J,1)] “deserving of respect or high regard”

  Эерэг .75 Сөрөг 0 Бодит .25

Туйлшралын үгийн сан хоорондын ялгаа

Christopher Potts, Sentiment Tutorial, 2011

	Үзэл бодлын үгийн сан	Ерөнхий лавлах	SentiWordNet	LIWC
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**	12/363 **(3%)**
Үзэл бодлын үгийн сан		32/2411 **(1%)**	1004/3994 **(25%)**	9/403 **(2%)**
Ерөнхий лавлах			520/2306 **(23%)**	1/204 **(0.5%)**
SentiWordNet				174/694 **(25%)**
LIWC				

IMDB дэх үг бүрийн туйлшралыг шинжлэвэл

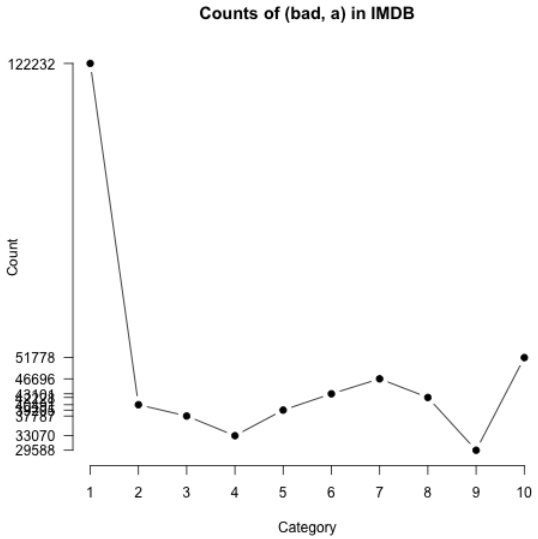
Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

- Хандлагын анги бүрд үг бүрийн магадлал хэр байдаг вэ?
- Тоог(“bad”) нь 1-од, 2-од, 3-од, гэх мэт үзвэл.
- Гэхдээ тохиолдох тоог ашиглаж болохгүй:
- Оронд нь, **магадлал:**

$$P(w \mid c) = \frac{f(w,c)}{\sum_{w \in c} f(w,c)}$$

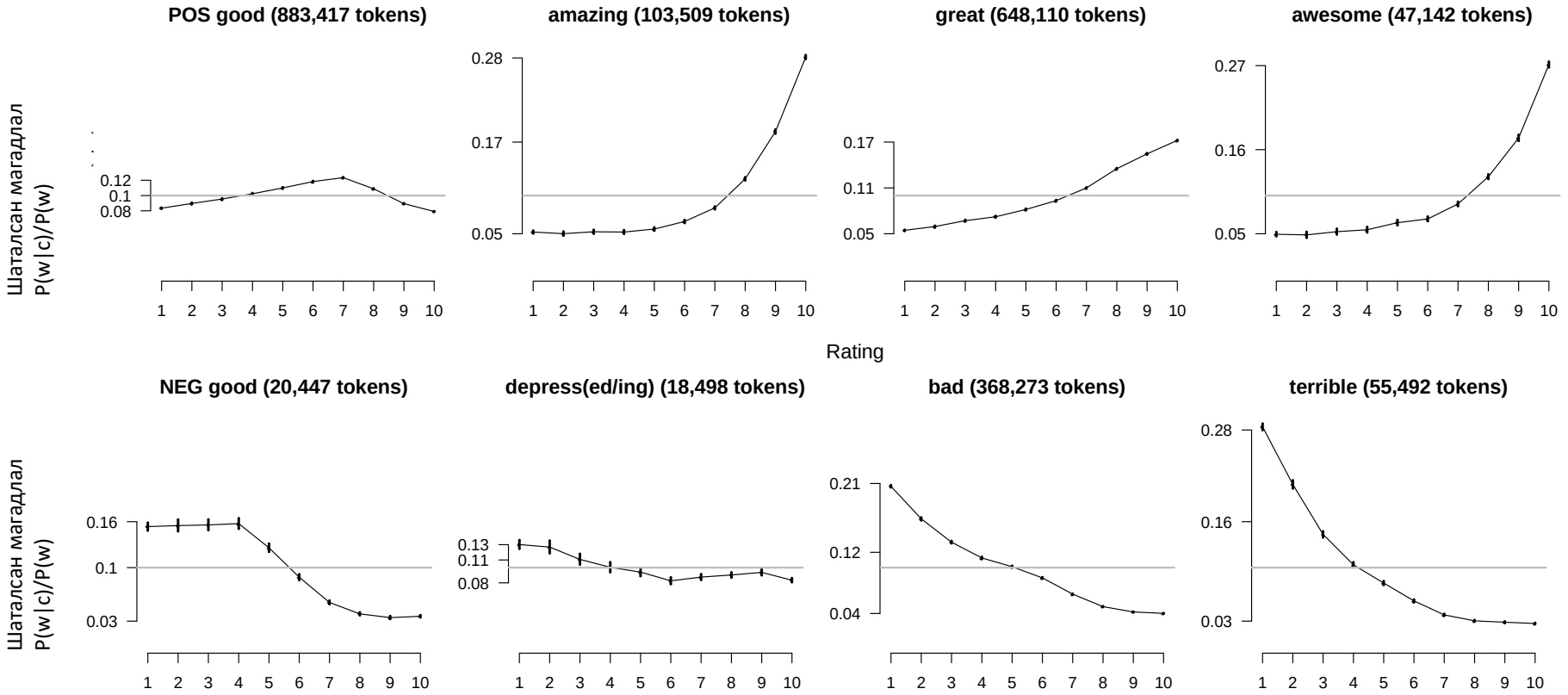
- үг хоорондын тэдгээрийг харьцуулж үзэх
  - **Шаталсан магадлал:**

$$\frac{P(w \mid c)}{P(w)}$$



IMDB дэх үг бүрийн туйлшралыг шинжлэвэл

Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

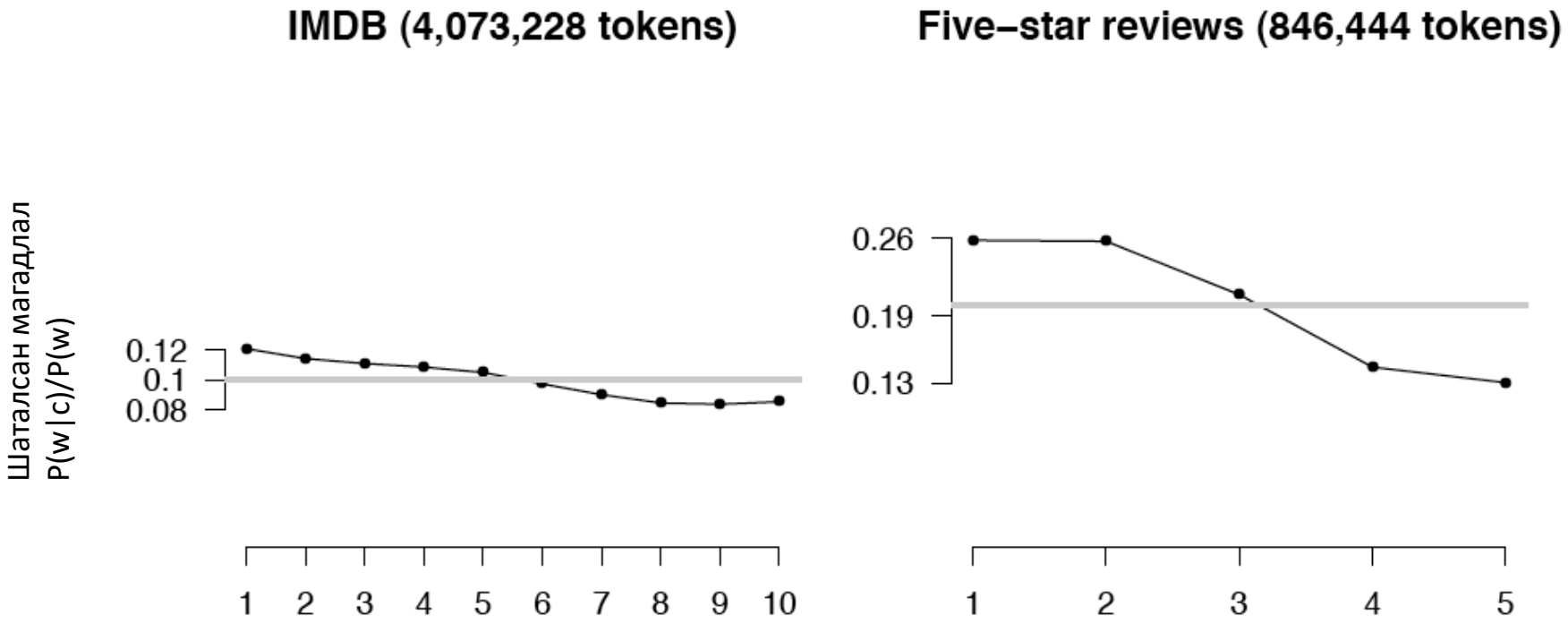


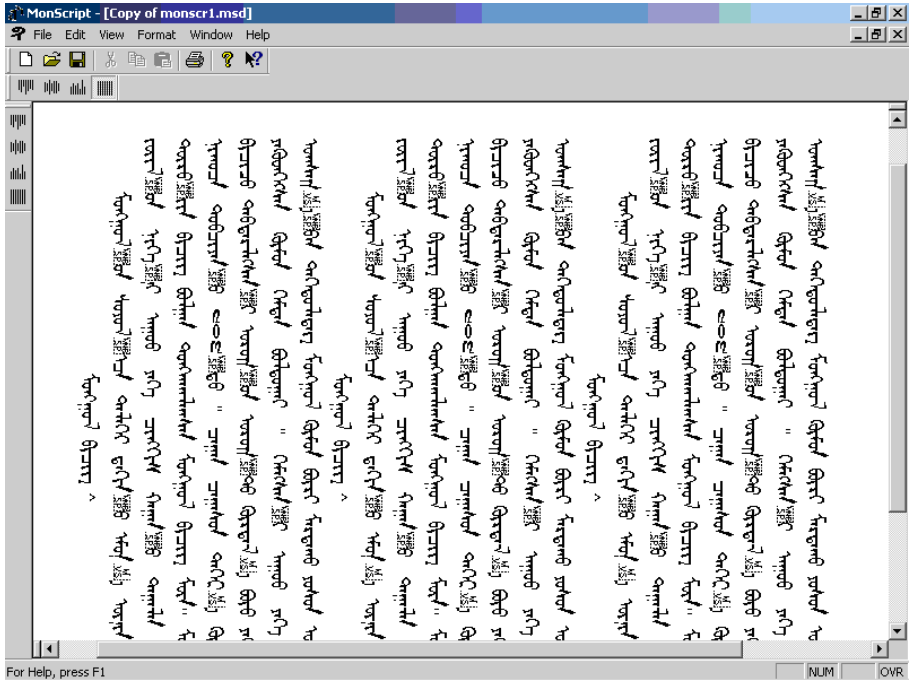
Бусад хандлагын онцлог шинж: Логик үгүйсгэл

Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

- логик үгүйсгэл (*no, not*) сөрөг хандлагатай холбоотой юу?
- Potts –ийн туршилт:
  - онлайн шүүмж дэх үгүйсгэлийг (*not, n't, no, never*) тоол
  - Шүүмжийн зэрэглэлийн эсрэг регресс

Potts 2011 үр дүн:
Сөрөг хандлага дахь олон үгүйсгэл





Хандлагын шинжилгээ

Хандлагын үгийн санг сургах

Үгийн сангийн хагас-supervised сургалт

- Жижиг хэмжээний мэдээлэл ашигла
  - Цөөн тэмдэглэсэн жишээнүүд
  - Цөөн гараар үүсгэсэн паттерн
- Үгийн санг bootstrap хийх

Hatzivassiloglou ба McKeown-ийн үгийн туйлшралыг таних санаа

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

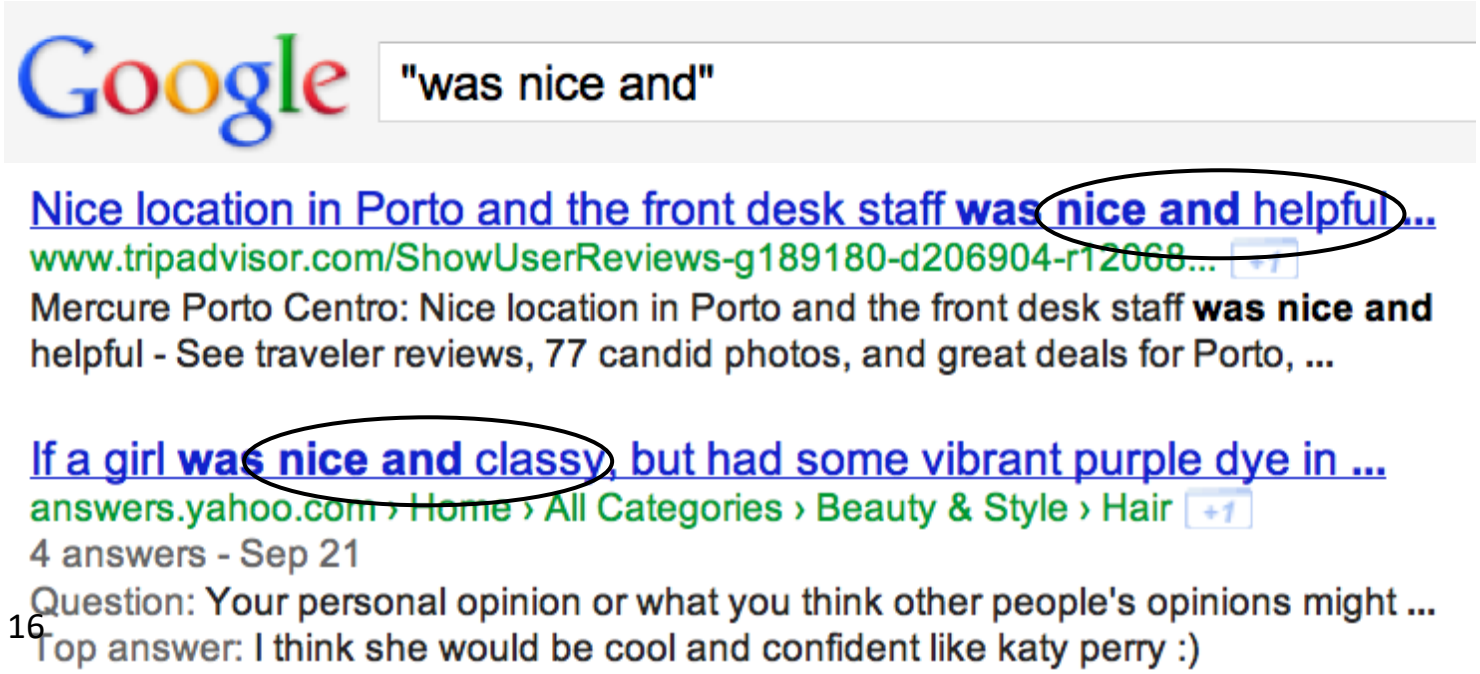
- “*and*” –ээр холбосон тэмдэг нэрүүд ижил туйлшралтай
  - Fair **and** legitimate, corrupt **and** brutal
  - *fair **and** brutal, *corrupt **and** legitimate
- “*but*” –аар холбогдсон тэмдэг нэрүүд эсрэг байна
  - fair **but** brutal

Hatzivassiloglou ба McKeown 1997
Алхам 1

- 1336 тэмдэг нэрийн **эхлэлийн олонлог**ийг тэмдэглэн (бүгд 21 сая үгтэй WSJ корпус дах 20 их)
  - 657 эерэг
    - adequate central clever famous intelligent remarkable reputed sensitive slender thriving…
  - 679 сөрөг
    - contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting…

Hatzivassiloglou ба McKeown 1997
Алхам 2

- Эхлэлийн олонлогийг холбосон тэмдэг нэрээр өргөтгөх

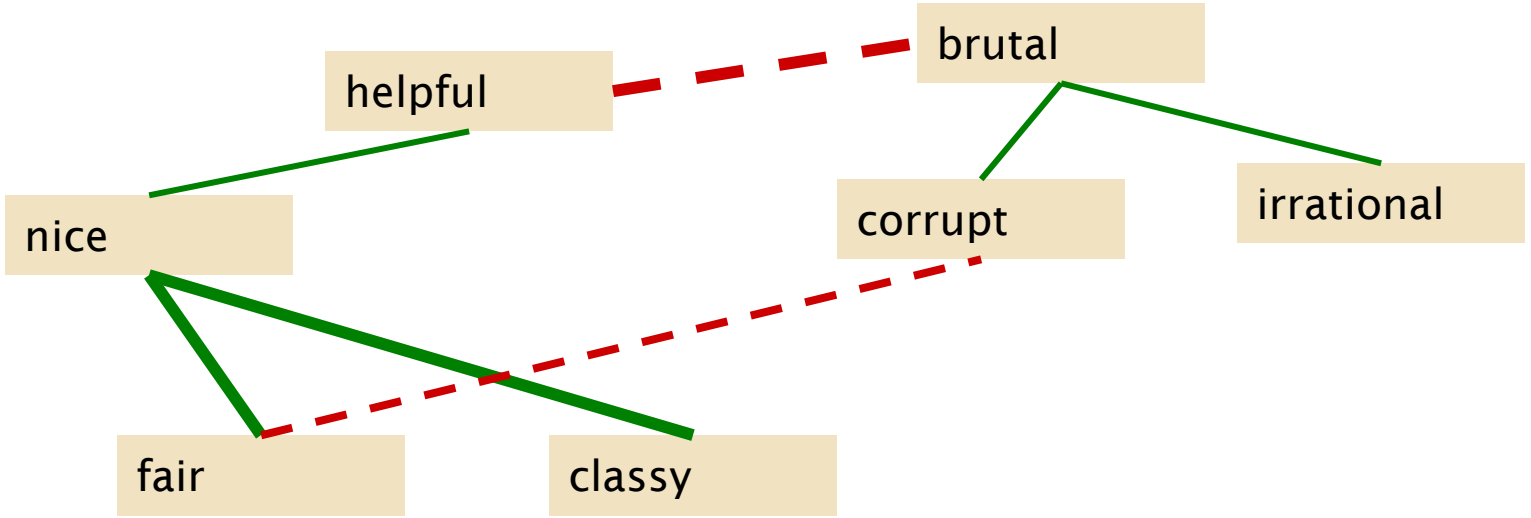


nice, helpful

nice, classy

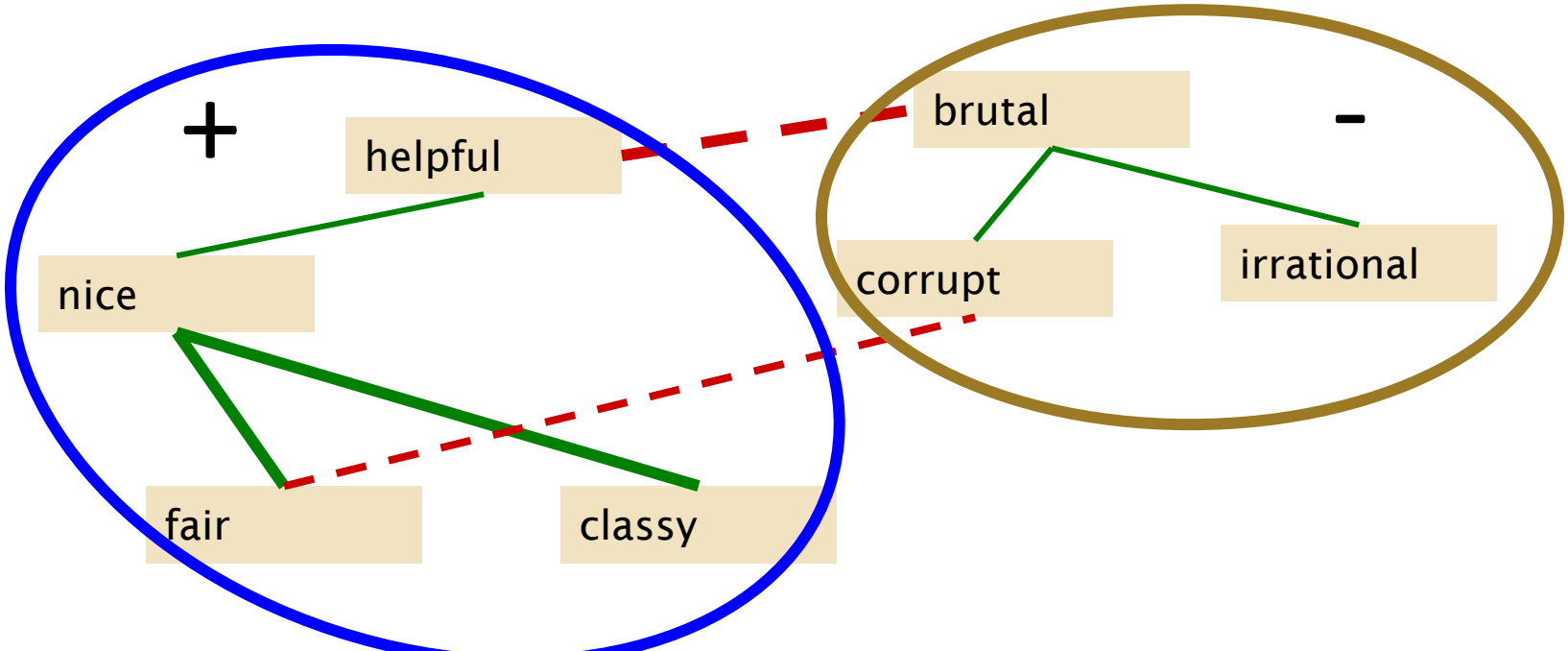
Hatzivassiloglou ба McKeown 1997 Алхам 3

- Supervised ангилагч нь үгийн хос бүрд “туйлшралын төсөөг” оноож, үр дүнг граф болгодог:



Hatzivassiloglou ба McKeown 1997
Алхам 4

- Графыг 2 хувааж кластер хийдэг



Туйлшралын үгийн санг гаргах

- Эерэг
  - bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty…
- Сөрөг
  - ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful…

Туйлшралын үгийн санг гаргах

- Эерэг
  - bold decisive **disturbing** generous good honest important large mature patient peaceful positive proud sound stimulating straightforward **strange** talented vigorous witty…
- Сөрөг
  - ambiguous **cautious** cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor **outspoken pleasant** reckless risky selfish tedious unsupported vulnerable wasteful…

Турнейн алгоритм

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

1. Шүүмжүүдээс хэлц үгсийг олборло
2. Хэл үг бүрийн туйлшралыг сурга
3. шүүмжид түүний хэлц үгсийн туйлшралын дундаар үнэлгээ өгөх

Тэмдэг нэр бүхий 2 үгтэй хэлц үгсийг олборлох

Эхний үг	Хоёр дах үг	3 дах үг (олборлоогүй)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

Хэлц үгийн туйлшралыг хэрхэн хэмжих вэ?

- Эерэг хэлц үг *“excellent”* үгтэй их хамт ордог
- Сөрөг хэл үг *“poor”* үгтэй их хамт ордог
- Гэвч хамт тохиолдохыг хэрхэн хэмжих вэ?

Pointwise Mutual Information

- 2 санамсаргүй X, Y хувсагчийн хоорондын **mutual information буюу харилцан хамаарал**

$$\mathrm{I}(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **Pointwise mutual information**:
  - хэрэв x болон y хараат бус байсан бол хамтдаа хэр их тохиолдох вэ?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Pointwise Mutual Information

- **Pointwise mutual information**:
  - хэрэв x болон y хараат бус байсан бол хамтдаа хэр их тохиолдох вэ?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **2 үг хоорондын PMI**:
  - хэрэв 2 үг хараат бус байсан бол хамтдаа хэр их тохиолдох вэ?

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

PMI –г хэрхэн тооцоолох вэ?

- Хайлтын хөдөлгүүр (Altavista) –аас асуух
  - P(word) нь `hits(word)/N` гэж үнэлэгдэнэ
  - P($word_1$,$word_2$) нь `hits(word1 NEAR word2)/N`^{2}

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{hits(word_1 \text{ NEAR } word_2)}{hits(word_1)hits(word_2)}$$

Хэлц үг “poor” эсвэл “excellent” -ийн альтай олон харагддаг вэ?

$$\text{Polarity}(phrase) = \text{PMI}(phrase, \text{"excellent"}) - \text{PMI}(phrase, \text{"poor"})$$

$$= \log_2 \frac{\text{hits}(phrase \text{ NEAR "excellent"})}{\text{hits}(phrase)\text{hits}(\text{"excellent"})} - \log_2 \frac{\text{hits}(phrase \text{ NEAR "poor"})}{\text{hits}(phrase)\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(phrase \text{ NEAR "excellent"})}{\text{hits}(phrase)\text{hits}(\text{"excellent"})} \frac{\text{hits}(phrase)\text{hits}(\text{"poor"})}{\text{hits}(phrase \text{ NEAR "poor"})}$$

$$= \log_2 \left(\frac{\text{hits}(phrase \text{ NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(phrase \text{ NEAR "poor"})\text{hits}(\text{"excellent"})} \right)$$

Сайшаасан шүүмжээс олборлосон хэлц үгс

Хэлц үг	POS tags	Туйлшрал
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
…		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
Average		0.32

Дургүйцсэн шүүмжээс олборлосон хэлц үгс

Хэлц үг	POS tags	Туйлшрал
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
…		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
Average		-1.2

Турнейн алгоритмын үр дүнгүүд

- Epinions –ийн 410 шүүмж
  - 170 (41%) сөрөг
  - 240 (59%) эерэг
- Дийлэнх анги нь: 59%
- Турнейн алгоритм: 74%

- Хэлц үг нь үгсээс илүү дээр
- Хэрэглээний хүрээний мэдээллээр сургана

Туйлшралыг сургахдаа WordNet ашиглах

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. COLING 2004
M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, 2004

- WordNet: онлайн тайлбар толь (өмнөх лекцид дурьдсан).
- Эерэг (“good”) ба сөрөг (“terrible”) эхлэлийн үгээр үүсгэнэ
- Ижил утгатай болон эсрэг утгатай үгсийг хайна
  - Эерэг олонлог: Эерэг (“well”) үгийн ижил утгатай үгс болон сөрөг үгсийн эсрэг утгатай үгсийг нэмнэ
  - Сөрөг олонлог: Сөрөг (“awful”) үгийн ижил утгатай үгс болон эерэг үгсийн эсрэг утгатай үгсийг нэмнэ
- Ижил утгатай үгсийн цувааг дагаж давтана
- Шүүнэ

Дүгнэлт: Үгийн санг сургах

- Давуу тал:
  - Хэрэглээний хүрээ зааж болно
  - Үгийг бодвол илүү найдвартай үр дүнтэй
- Санаа Intuition
  - (‘good’, ‘poor’) гэсэн эсрэг эхлэлийн үгсээс эхэлнэ
  - Ижил туйлшралтай бусад үгсийг хайна:
    - “and” болон “but” ашиглана
    - Нэг документ дээр ойролцоо гарч байгаа үгсийг ашиглах
    - WordNet дэх ижил болон усрэг утгатай үгсийг ашиглах