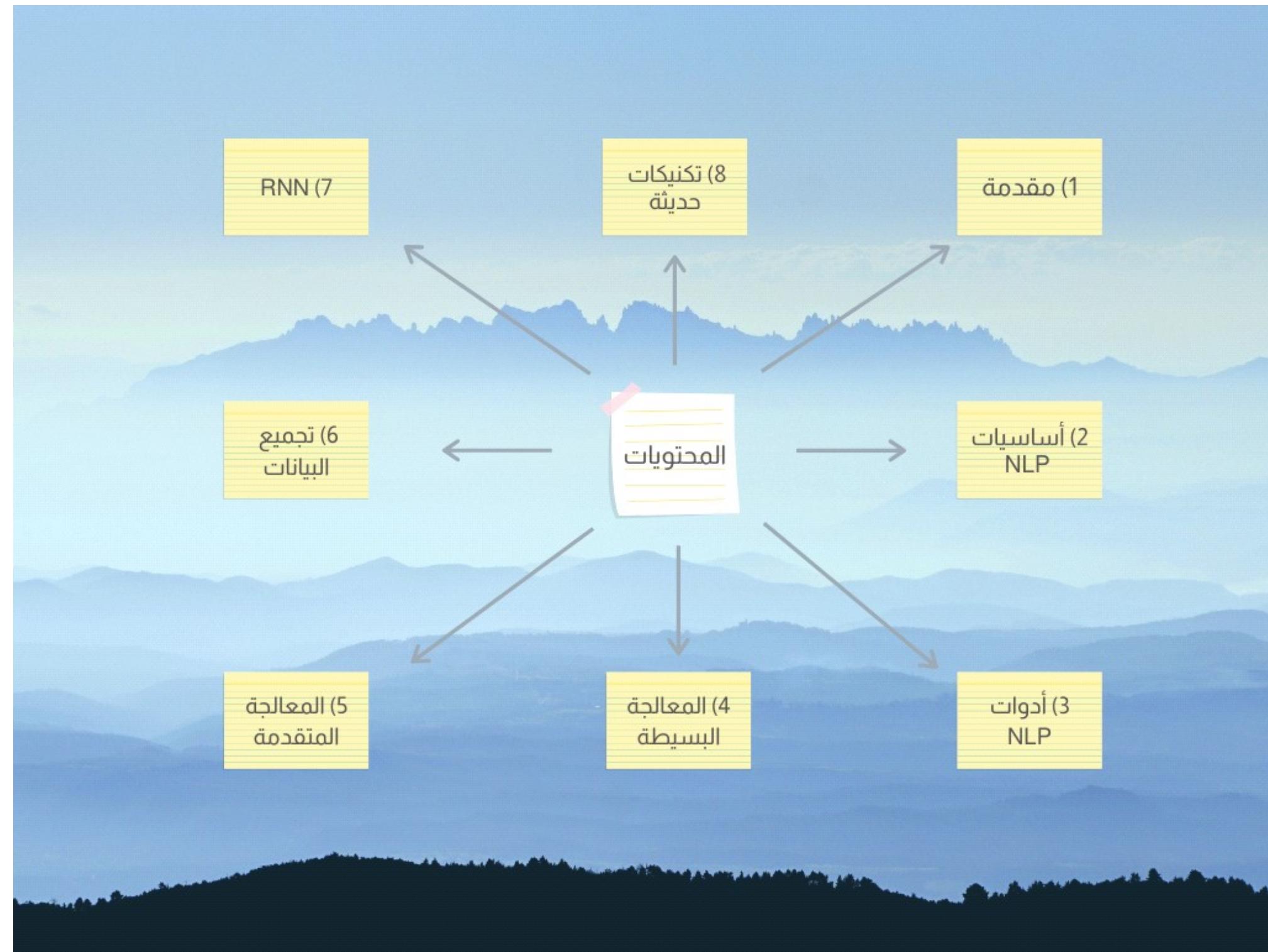


# NATURAL LANGUAGE PROCESSING

# المعالجة اللغوية الطبيعية



# المحتويات

					التطبيقات	العقبات و التحديات	NLP تاريخ ملفات pdf	ما هو NLP الملفات النصية	المحتويات المكتبات	1) مقدمة
T.Visualization	Syntactic Struc.	Matchers	Stopwords	NER	Stem & Lemm	POS	Sent. Segm.	Tokenization	NLP	2) أساسيات NLP
	Dist. Similarity	Text Similarity	TF-IDF	BOW	Word2Vec	T. Vectors	Word embed	Word Meaning		3) أدوات NLP
T. Generation	L. Modeling	NGrams	Lexicons	GloVe	NMF	LDA	T. Clustering	T. Classification		4) المعالجة البسيطة
	Summarization & Snippets		Ans. Questions		Auto Correct	Vader	Naïve Bayes	Sent. Analysis		5) المعالجة المتقدمة
Search Engine	Relative Extraction		Information Retrieval		Information Extraction		Data Scraping	Tweet Collecting		6) تجميع البيانات
					Rec NN\TNN	GRU	LSTM	Seq to Seq		7) RNN (
Chat Bot	Gensim	FastText	Bert	Transformer	Attention Model	T. Forcing	CNN	Word Cloud		8) تكنيات حديثة

## القسم الخامس : المعالجة المتقدمة للنصوص

### الجزء السابع : المعاجم Lexicons

فكرة المعاجم أو **lexicons** , هي قائمة على معرفة معاني عدد من الكلمات الهامة التي سيتم استخدامها بشكل مكثف ( الاف او عشرات الالاف ) , وتحديد معناها و مدى قوتها في عوامل معينة , وبالتالي يكون التصنيف قائما على أساس دقة . . .

و هناك نوعين أساسيين من المعاجم :

- المعجم المصنف , وهو الذي فقط هناك عدد من الكلمات الايجابية في ملف , وكلمات اخرى سلبية في ملف آخر , أي أن يحتوي المعجم على عشرات الالاف من الكلمات الايجابية و السلبية , مع عنونة هذه الكلمات , لمعرفة ان هذه كلمات ايجابية و سلبية .
- المعجم المفسر . و بعض المعاجم لا تحتوي فقط على تقسيم, بل انه يكون اشبه بالقاموس لأنه يحتوي على معاني الكلمات , والمفهوم العام عن كل كلمة , مع وضع في الاعتبار ان ال **lexicon** من الممكن ان يكون عام او بتخصص معين , فكلمة "شيخ" يتم فهمها في اللهجة المصرية بشكل مختلف عن اللهجة السعودية , و كذلك كلمة "هدف" يتم فهمها في المقال الرياضي , بشكل مختلف عن المقال الاقتصادي

ومن أهم استخداماتها :

- Sentimental Analysis
- استخراج المعلومات
- الإجابة عن الأسئلة
- الترجمة الآلية

و هناك عدة نماذج منه ، فجامعة هارفارد أصدرت نسخة فيها 1915 كلمة ايجابية و 2291 كلمة سلبية ، وفيها تفاصيل كاملة عن كل كلمة ، و هنا الروابط

<http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls>

<http://www.wjh.harvard.edu/~inquirer/homecat.htm>

<http://www.wjh.harvard.edu/~inquirer/Positiv.html>

<http://www.wjh.harvard.edu/~inquirer/Negativ.html>

كما أن عددا من الباحثين بولاية تكساس أصدروا قاعدة بيانات LIWC ، وفيها عدد كبير من المعاني



## LIWC (Linguistic Inquiry and Word Count)

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

- Home page: <http://www.liwc.net/>
- 2300 words, >70 classes
- **Affective Processes**
  - negative emotion (*bad, weird, hate, problem, tough*)
  - positive emotion (*love, nice, sweet*)
- **Cognitive Processes**
  - Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)
- **Pronouns, Negation (*no, never*), Quantifiers (*few, many*)**
- \$30 or \$90 fee

أيضا معجم MPQA ، والتي فيها 8221 جذر للكلمة ، و كذلك 6885 كلمة ، منها 2718 كلمة ايجابية و 4912 كلمة سلبية ، و التي هنا :

[http://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon/](http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/)

أيضا معجم Bing Liu و الذي به 2006 كلمة ايجابية ، و 4783 كلمة سلبية الذي هنا

<https://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>

كما ان عددا من الباحثين في ايطاليا قاموا بترتيب الكلمات على مقياس : ايجابي ، سلبي ، متعادل

<https://www.aclweb.org/anthology/L10-1531/>

All WordNet synsets automatically annotated for degrees of positivity,  
negativity, and neutrality/objectiveness

[estimable(J,3)] “may be computed or estimated”

Pos 0 Neg 0 Obj 1

[estimable(J,1)] “deserving of respect or high regard”

Pos .75 Neg 0 Obj .25

و نري ان هذه المعاجم و الخوارزميات ليست متطابقة تماما ، لكن مساحة التشابه كبيرة

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
General Inquirer			520/2306 (23%)	1/204 (0.5%)
SentiWordNet				174/694 (25%)
LIWC				

كما أنهم قاموا بحساب عدد بعض الكلمات في تقييمات الافلام الايجابية و السلبية , لمعرفة مدى تكرار كلمة مثل bad في الافلام ذات نجمة واحدة او نجمتين الى 10 نجوم , و كأنها مؤشر لمدى معنى الكلمة

How likely is each word to appear in each sentiment class?

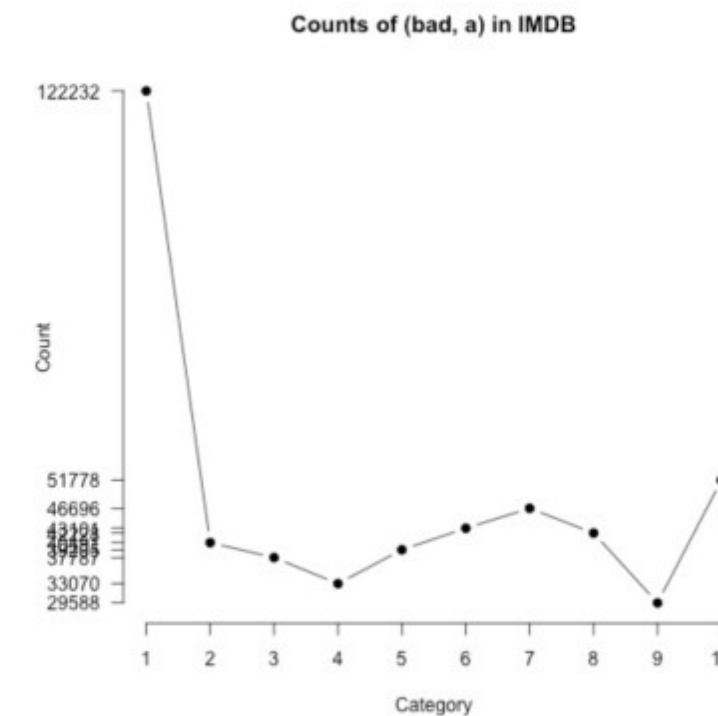
Count("bad") in 1-star, 2-star, 3-star, etc.

But can't use raw counts:

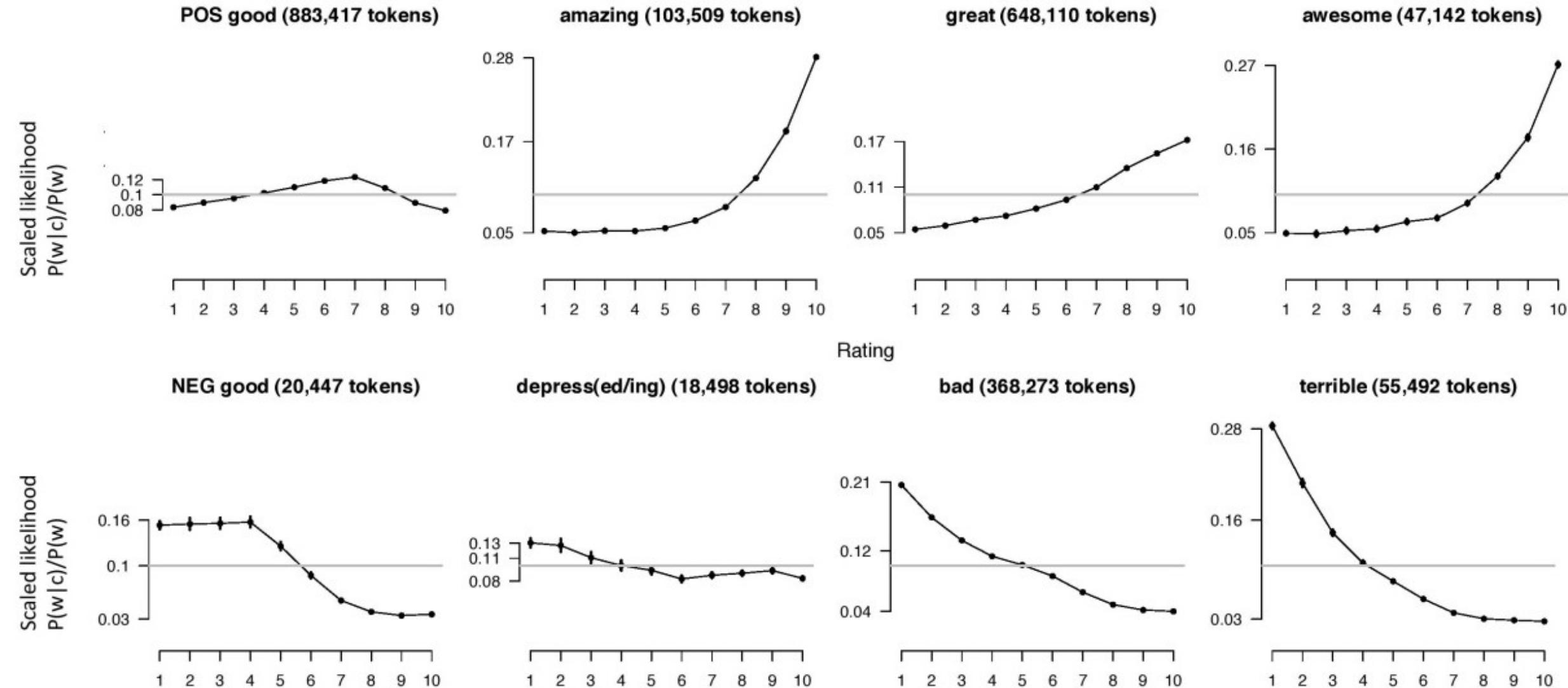
Instead, **likelihood**:  $P(w|c) = \frac{f(w,c)}{\sum_{w \in c} f(w,c)}$

Make them comparable between words

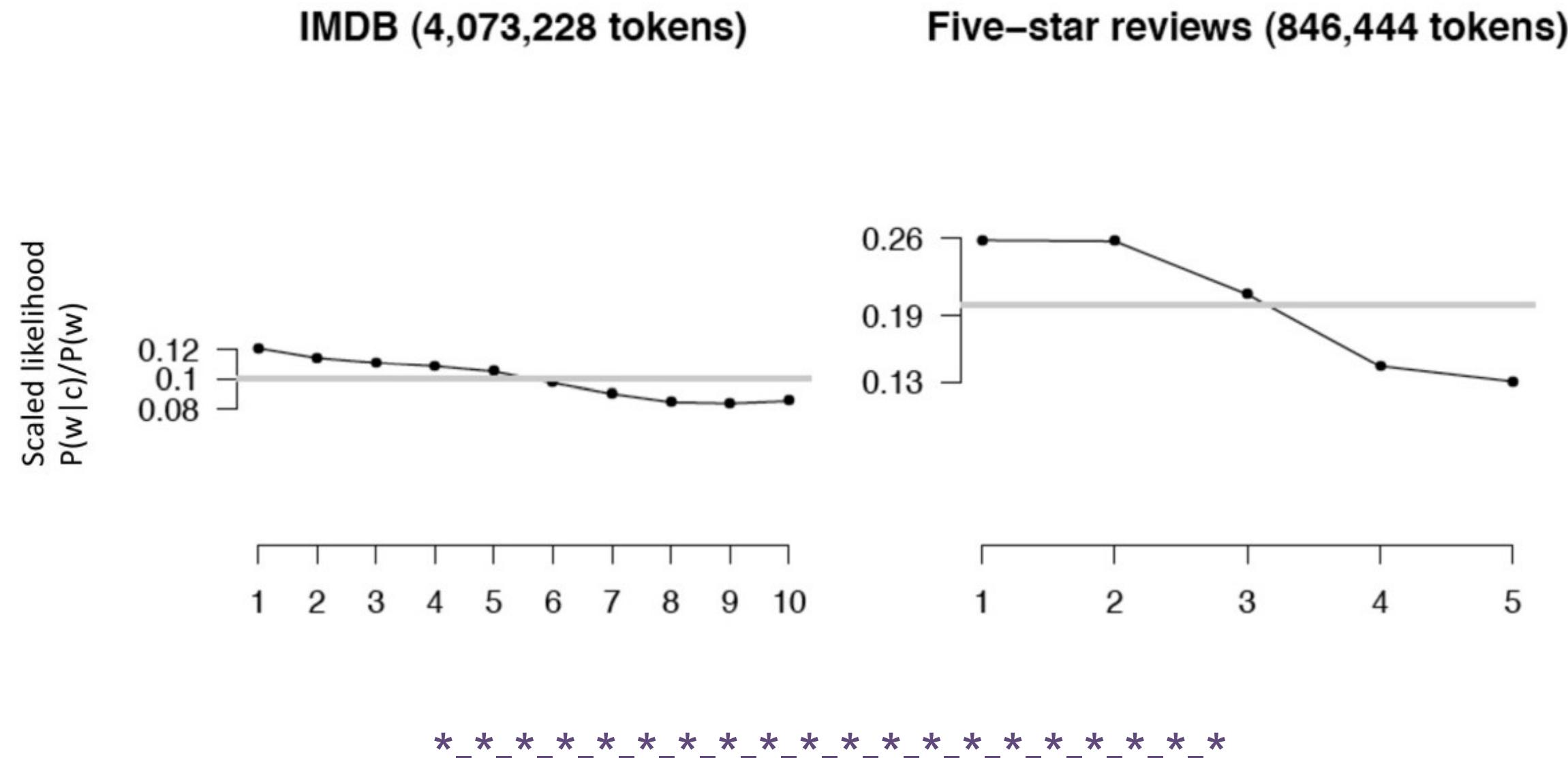
- **Scaled likelihood**:  $\frac{P(w|c)}{P(w)}$



# ونري تطبيق هذا الامر على عدد من الكلمات في العينة الايجابية و السلبية



اًيضاً قاموا بالتركيز على كلمات النفي (no , not , never , n't) لحساب عددها في العينات السلبية و الايجابية ، فكان هناك تأثيراً طفيفاً لها



نتكلم الان عن تكوين و تدريب هذا النوع من الخوارزميات ، كيف يقومون بمعارفه معاني كلمات الاف الكلمات مثلما رأينا في الخوارزميات السابقة ؟ ؟

غالباً ما يتم هذا عبر استخدام التدريب بنصف إشراف semi-supervised ML ، أي أن يتم عمل عينة محدودة من العناصر ، ثم استخدامها كبيانات معونة لتدريب موديل و عمل عناوين لباقي العينة

فمثلاً معرفة ان الكلمة **fair** لها معانٍ معينة ، ثم البحث عن الصفات adjectives التي جاءت تالية لها بعد الكلمة **ad** فنعرف انها قريبة منها في المعنى

او البحث عن الكلمات التي أتت مع الكلمة **but** فنعرف انها عكسها في المعنى

**Adjectives conjoined by “and” have same polarity**

- **Fair and legitimate, corrupt and brutal**
- **\*fair and brutal, \*corrupt and legitimate**

**Adjectives conjoined by “but” do not**

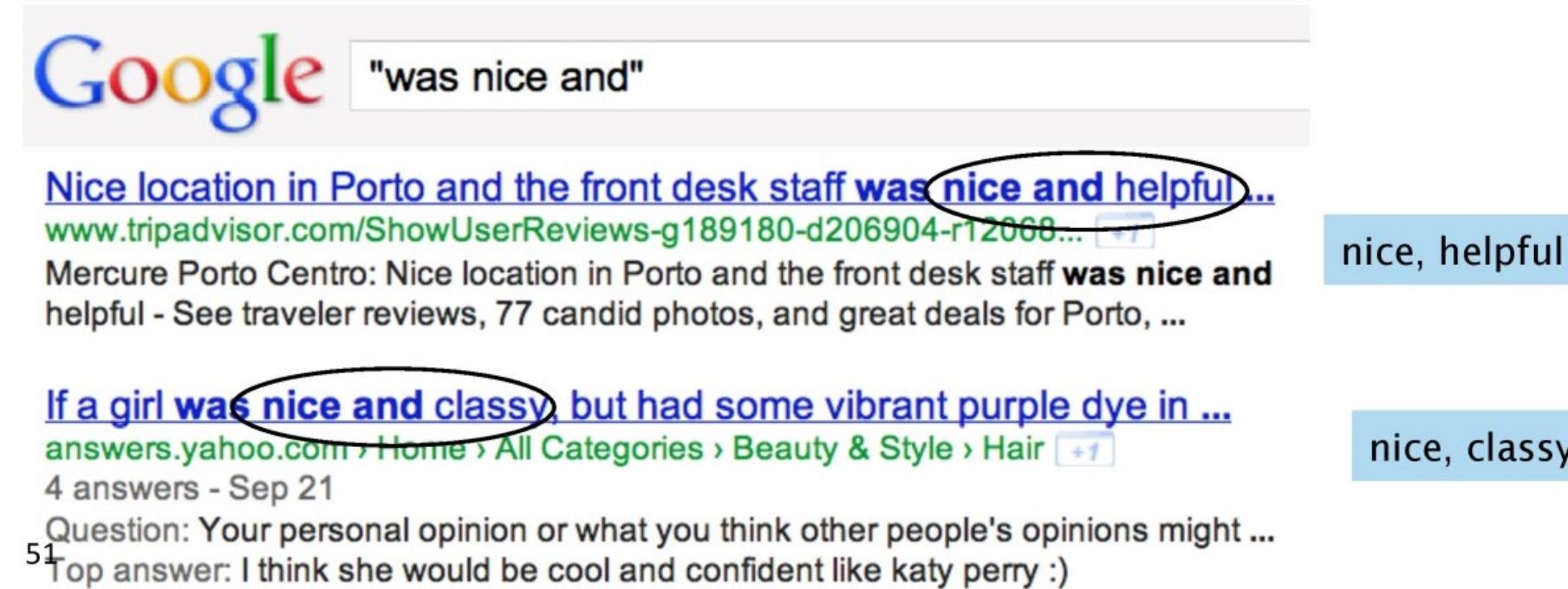
- **fair but brutal**

و قد قام هذان الباحثان بعمل معجم مشابه ، يحتوي علي حوالي 600 معني ايجابي و 600 سلبي هكذا .

## Label seed set of 1336 adjectives (all >20 in 21 million word WSJ corpus)

- 657 positive
  - adequate central clever famous intelligent remarkable reputed sensitive slender thriving...
- 679 negative
  - contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting...

ثم قاموا بالبحث في جوجل علي الكلمات المشابهة لها ، لمعرفة الكلمات المرادفة لها ، لا تنس أن جوجل يأتي بالكلمات المشابهة من المقالات و المواقع ، وكأنه يستخدم جميع البيانات الموجودة اونلاين للبحث فيها عن كلمات مشابهة

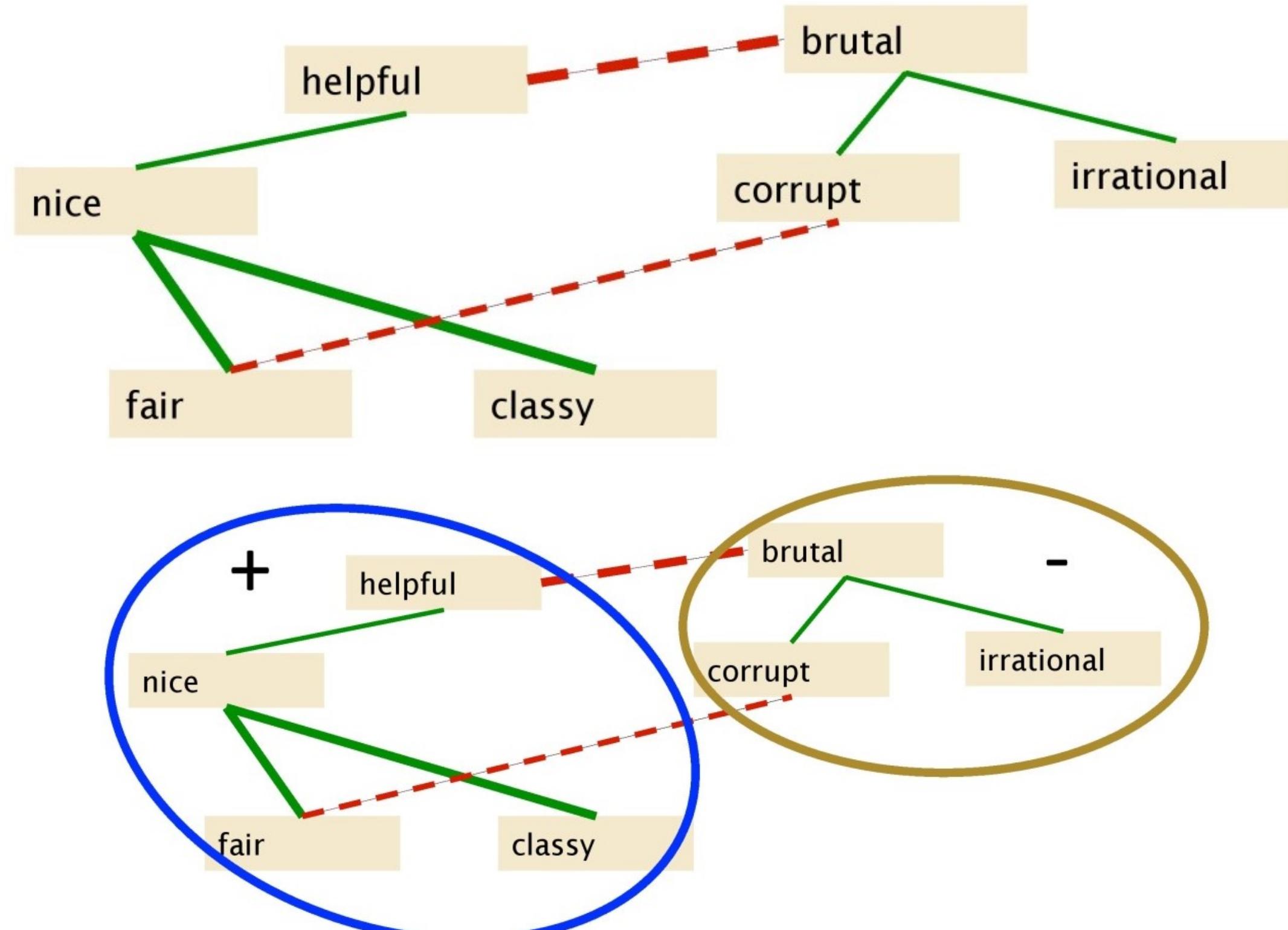


Google search results for "was nice and".

The search results show two examples:

- Nice location in Porto and the front desk staff was nice and helpful...** (highlighted with a blue oval)  
www.tripadvisor.com>ShowUserReviews-g189180-d206904-r12008...  
Mercure Porto Centro: Nice location in Porto and the front desk staff was nice and helpful - See traveler reviews, 77 candid photos, and great deals for Porto, ...  
A blue box labeled "nice, helpful" is positioned next to this result.
- If a girl was nice and classy, but had some vibrant purple dye in ...** (highlighted with a blue oval)  
answers.yahoo.com › Home › All Categories › Beauty & Style › Hair +1  
4 answers - Sep 21  
Question: Your personal opinion or what you think other people's opinions might ...  
Top answer: I think she would be cool and confident like katy perry :)

ثم يتم الربط بين الكلمات القريبة و البعيدة عن بعضها البعض , ثم القيام بعمل عناقيد clustering لتحديد نوعية كل عنقود من الكلمات



لكن للأمر عيوبه ، فهناك بعض الكلمات ذات المعنى الايجابي موجودة في العينة السلبية و العكس

## Positive

- bold decisive **disturbing** generous good honest important large mature patient peaceful positive proud sound stimulating straightforward **strange** talented vigorous witty...

## Negative

- ambiguous **cautious** cynical evasive harmful hypocritical inefficient  
insecure irrational irresponsible minor **outspoken pleasant** reckless risky  
selfish tedious unsupported vulnerable wasteful...

كما ان هناك خوارزم آخر يسمى Turney و الذي يقوم على خطوات ثلاث :

- استخراج جملة من التعليقات
  - تحديد مدى قطبية كل جملة
  - حساب درجة الجملة بناء على متوسط الجمل

و يقصد بالقطبية هنا مدي اقترابها من المعنى الايجابي او السلبي ، فجملة "I love it" لها اقتراب كبير مع المعنى الايجابي ، و جملة "it's scrap" لها اقتراب كبير مع المعنى السلبي

و لدي تطبيق فكرة استخراج الجمل ، قام الباحث بالاعتماد على نوعيات معينة من الجمل مثل هذه :

- Adjective + Noun ( or Plural Noun ) :
  - Awesome car works quickly
- Adverb + Adjective
  - I am extremely happy in my new job.
- Adjective + Adjective :
  - I have a great big book
- Noun + Adjective :
  - Is car quick enough ? ?
- Adverb + verb :
  - She quickly agreed to this point

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

و من هنا نقوم بحساب القطبية حسب اقتراب الجملة او ابعادها عن المعنى الايجابي او السلبي

Positive phrases co-occur more with “*excellent*”

Negative phrases co-occur more with “*poor*”

But how to measure co-occurrence?

و لكن كيف نقوم بحساب الاقتراب و الابتعاد ؟

هذا القانون لحساب ما يسمى " نقطة تبادل المعلومات " , اي المساحات المشتركة بين كلمتين او جملتين .. و يكون هذا عبر حساب لوغاريتم احتمالية تواجد الكلمتين معا ( لصيقتين و قريبتين من بعضهما البعض ) , مقسومة على احتمال تواجد الكلمة الاولى مضروبة في احتمال الكلمة الثانية

### **Pointwise mutual information:**

- How much more do events x and y co-occur than if they were independent?

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

### **PMI between two words:**

- How much more do two words co-occur than if they were independent?

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$$

و حساب الاحتمالية يكون عبر حساب تكرار الكلمة في عينة البيانات , مقسوما على عدد الكلمات الكلي

بينما حساب احتمالية تواجد كلمتين قريبتين معا , هو ضرب عدد مرات تكرار الكلمة الاولى في عدد مرات تكرار الثانية مقسوما على مربع عدد كلمات العينة

لذا يكون القانون هكذا

- Query search engine (Altavista)
  - $P(\text{word})$  estimated by  $\text{hits}(\text{word})/N$
  - $P(\text{word}_1, \text{word}_2)$  by  $\text{hits}(\text{word}_1 \text{ NEAR } \text{word}_2)/N^2$

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{\text{hits}(\text{word}_1 \text{ NEAR } \text{word}_2)}{\text{hits}(\text{word}_1)\text{hits}(\text{word}_2)}$$

و يمكن تطبيق هذا الامر على جملة ما , لحساب مدى اقتراب الجملة من معنی ايجابي , او معنی سلبي , لحساب درجة او تقييم لها

مع العلم ان الجمل يمكن ان تكون صغيرة , و اغلبها هي كلمتين متتاليتين فقط , ولكن بأحد النماذج السابق شرحها , لأن النماذج الأخرى قد لا تكون مفيدة لنا في التدريب



## Does phrase appear more with “poor” or “excellent”?

Polarity(*phrase*) = PMI(*phrase*, "excellent") – PMI(*phrase*, "poor")

$$\begin{aligned} &= \log_2 \frac{\text{hits}(\textit{phrase NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}("excellent")} - \log_2 \frac{\text{hits}(\textit{phrase NEAR "poor"})}{\text{hits}(\textit{phrase})\text{hits}("poor")} \\ &= \log_2 \frac{\text{hits}(\textit{phrase NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}("excellent")} \frac{\text{hits}(\textit{phrase})\text{hits}("poor")}{\text{hits}(\textit{phrase NEAR "poor"})} \\ &= \log_2 \left( \frac{\text{hits}(\textit{phrase NEAR "excellent"})\text{hits}("poor")}{\text{hits}(\textit{phrase NEAR "poor"})\text{hits}("excellent")} \right) \end{aligned}$$

62

و بالتالي اذا قمنا بحساب درجات الجمل في عدد كبير من التعليقات الايجابية عن خدمات بنكية مثلا ، ستكون مدي اقتراب كل جملة من المعنى الايجابي هكذا

# Phrases from a thumbs-down review

Phrase	POS tags	Polarity
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
<i>Average</i>		-1.2

# Phrases from a thumbs-up review

Phrase	POS tags	Polarity
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
<i>Average</i>		0.32

هنا نتناول الحديث عن قاموس word net 3.0 و الذي اصدره طلاب جامعة برنستون بولاية نيو جيرسي

و هو قاموس لعدد كبير من الكلمات الانجليزية ، ويتم التحضير لعدد اخر من اللغات ، ويمكن تحميله من مكتبة nltk

او استخدامه من هنا

<http://wordnetweb.princeton.edu/perl/webwn>

او تحميله من هنا

<https://wordnet.princeton.edu/download/current-version>

كما يمكن تحميله من مكتبة nltk

و يتكون القاموس من :

Category	Unique Strings
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

و يكون شكله كالتالي : لفظ bass

#### Noun

- S: (n) bass (the lowest part of the musical range)
- S: (n) bass, bass part (the lowest part in polyphonic music)
- S: (n) bass, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) bass, bass voice, basso (the lowest adult male singing voice)
- S: (n) bass (the member with the lowest range of a family of musical instruments)
- S: (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

#### Adjective

- S: (adj) bass, deep (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

و هنا تعريف لفظ sense

The synset (synonym set), the set of near-synonyms, instantiates a sense or concept, with a gloss

Example: chump as a noun with the gloss:

"a person who is gullible and easy to take advantage of"

This sense of "chump" is shared by 9 words:

chump<sup>1</sup>, fool<sup>2</sup>, gull<sup>1</sup>, mark<sup>9</sup>, patsy<sup>1</sup>, fall guy<sup>1</sup>, sucker<sup>1</sup>, soft touch<sup>1</sup>, mug<sup>2</sup>

Each of these senses have this same gloss

- (Not every sense; sense 2 of gull is the aquatic bird)

## و هنا تفاصيل اخرى فيها

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> <sup>1</sup> → <i>meal</i> <sup>1</sup>
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> <sup>1</sup> → <i>lunch</i> <sup>1</sup>
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> <sup>2</sup> → <i>professor</i> <sup>1</sup>
Has-Instance		From concepts to instances of the concept	<i>composer</i> <sup>1</sup> → <i>Bach</i> <sup>1</sup>
Instance		From instances to their concepts	<i>Austen</i> <sup>1</sup> → <i>author</i> <sup>1</sup>
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> <sup>1</sup> → <i>crew</i> <sup>1</sup>
Part Meronym	Has-Part	From wholes to parts	<i>table</i> <sup>2</sup> → <i>leg</i> <sup>3</sup>
Part Holonym	Part-Of	From parts to wholes	<i>course</i> <sup>7</sup> → <i>meal</i> <sup>1</sup>
Antonym		Opposites	<i>leader</i> <sup>1</sup> → <i>follower</i> <sup>1</sup>

## MeSH (Medical Subject Headings)

- 177,000 entry terms that correspond to 26,142 biomedical “headings”

### Hemoglobins

Synset

**Entry Terms:** Eryhem, Ferrous Hemoglobin, Hemoglobin

**Definition:** The oxygen-carrying proteins of ERYTHROCYTES.

They are found in all vertebrates and some invertebrates.

The number of globin subunits in the hemoglobin quaternary structure differs between species. Structures range from monomeric to a variety of multimeric arrangements

1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. - Chemicals and Drugs [D]
  - [Inorganic Chemicals \[D01\]](#) +
  - [Organic Chemicals \[D02\]](#) +
  - [Heterocyclic Compounds \[D03\]](#) +
  - [Polycyclic Compounds \[D04\]](#) +
  - [Macromolecular Substances \[D05\]](#) +
  - [Hormones, Hormone Substitutes, and Hormone Antagonists \[D06\]](#) +
  - [Enzymes and Coenzymes \[D08\]](#) +
  - [Carbohydrates \[D09\]](#) +
  - [Lipids \[D10\]](#) +
  - [Amino Acids, Peptides, and Proteins \[D11\]](#) +
  - [Nucleic Acids, Nucleotides, and Nucleosides \[D12\]](#) +
  - [Complex Mixtures \[D20\]](#) +
  - [Biological Factors \[D23\]](#) +
  - [Biomedical and Dental Materials \[D25\]](#) +

[Proteins \[D12.776\]](#)

[Blood Proteins \[D12.776.124\]](#)

[Acute-Phase Proteins \[D12.776.124.050\]](#) +

[Anion Exchange Protein 1, Erythrocyte \[D12.776.124.078\]](#)

[Ankyrins \[D12.776.124.080\]](#)

[beta 2-Glycoprotein I \[D12.776.124.117\]](#)

[Blood Coagulation Factors \[D12.776.124.125\]](#) +

[Cholesterol Ester Transfer Proteins \[D12.776.124.197\]](#)

[Fibrin \[D12.776.124.270\]](#) +

[Glycophorin \[D12.776.124.300\]](#)

[Hemocyanin \[D12.776.124.337\]](#)

► [Hemoglobins \[D12.776.124.400\]](#)

[Carboxyhemoglobin \[D12.776.124.400.141\]](#)

[Erythrocytins \[D12.776.124.400.220\]](#)



و يمكن استخدام wordnet من مكتبة nltk , والتي تقوم بتناول الكلمة و اعطاء التفسير الكامل لها , و يتم استدعائها عبر الدالة synsets التعرف على المعاني المختلفة للكلمة , و اعطاء امثلة عليها

و هناك عدد من الأوامر الهامة فيها مثل :

- wordnet.synsets(word) : اعطاء المعاني المختلفة للكلمة
- .name() : اسم هذا المعنى
- .definition() : تعرف كامل لهذا المعنى
- .examples() : مثال على المعنى
- .antonyms() : الكلمات المضادة
- w1.wup\_similarity(w2) : مدى اقتراب الكلمة الاولى من الثانية

\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*-\*