

# Opinion Mining

(tratto in parte da “Opinion Mining”, Bing Liu  
e “Opinion Mining and Sentiment Analysis” B.  
Pang & L. Lee)

## Fatti e Opinioni

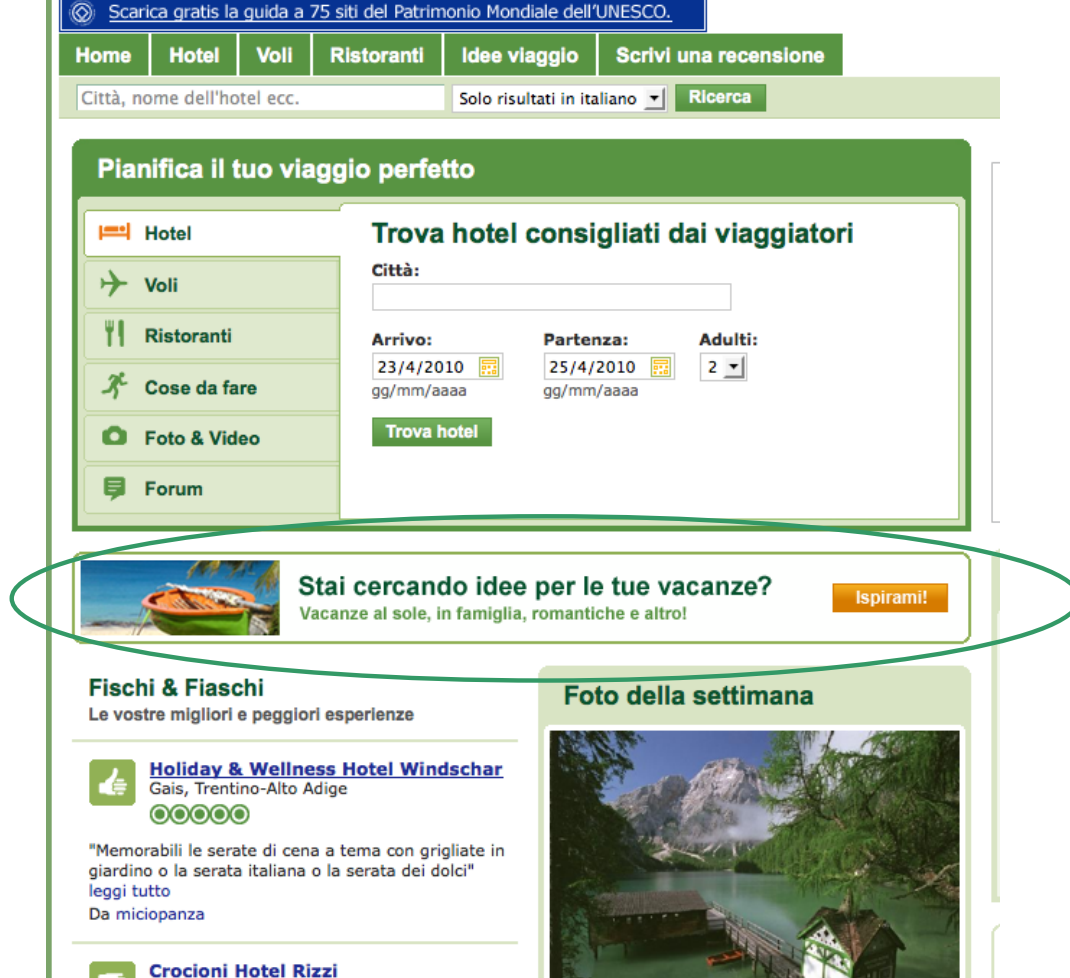
- Nel web sono memorizzati due tipi di informazioni testuali: **fatti** e **opinioni**
- I motori di ricerca ricercano **fatti**, mediante keywords relative all'argomento della ricerca
- Le opinioni non possono essere ricercate mediante parole chiave (*“cosa pensa la gente dei cellulari Nokia?”*)
- Le attuali strategie di ranking NON sono adatte al recupero di opinioni

# Opinioni: contenuto generato dagli utenti

- Gli utenti del web esprimono opinioni su qualsiasi argomento (servizi, prodotti, politica..). Queste informazioni vengono classificate come “**user generated content**”
- Recentemente questo tipo di contenuti è diventato dimensionalmente molto rilevante
- Da un punto di vista scientifico e applicativo, è molto interessante poter analizzare le opinioni

## Applicazioni

- **Business e organizzazioni:** benchmarking di prodotti e servizi
  - Si spende molto denaro per analizzare opinioni e sentimenti di utenti
- **Singoli utenti:** sono interessati alle opinioni altrui quando
  - Devono acquistare un prodotto o servizio
  - Sono interessati alle altrui opinioni su temi politici o di attualità
- **Pubblicità:** per piazzare un *advertisement* dove gli utenti generano contenuto



## Due tipi di valutazioni

- **Opinioni soggettive**: valutazioni su oggetti, prodotti, eventi, argomenti..
  - “the picture quality of this camera is great”
- **Paragoni (oggettivi o soggettivi)**: espressioni che evidenziano similarità e differenze fra diversi oggetti, in genere definendo un ordinamento
  - “car x is cheaper than car y.”

# Opinion Search

- E' possibile cercare opinioni con la stessa efficienza con cui si cercano fatti?
- Non ancora.. (ma è molto interessante poterlo fare)
- **Google Opinion** coming soon?

## Opinion Queries: tipi

1. Trovare l'opinione (articolata) di una persona o organizzazione (*opinion holder*) su un particolare oggetto o caratteristica di oggetto (*feature*)  
*“cosa pensa Obama dell'aborto?”*
2. Trovare opinioni di vari opinion holders su vari oggetti/features
  - positive o negative  
*Cosa pensano i turisti degli alberghi romani?*
  - Per analizzare l'evolversi di opinioni nel tempo  
*Le minicar raccolgono maggiori o minori consensi?*
  - per comparare due oggetti A e B  
*Gmail o Hotmail?*

# 1. Trovare l'opinione di una persona

- Per questo tipo di ricerca, le tradizionali search engine riescono a recuperare informazioni utili
- Perché:
  - Un opinion holder ha in genere UNA opinione su UN topic
  - Questa opinione in genere è completamente espressa in un singolo documento
  - Quindi, delle keywords appropriate (Obama, aborto) possono essere appropriate



The screenshot shows a Google search interface with the query 'obama aborto' entered in the search bar. Below the search bar, there are radio buttons for 'Cerca: nel Web', 'Cerca: pagine in Italiano', and 'Cerca: pagine provenienti da: Italia'. The search results are displayed under the 'Web' tab, with a link to 'Mostra opzioni...'. The results list several articles related to Obama's stance on abortion, including dates, headlines, and snippets of text. Each result includes a link to the full article and options for 'Copia cache' and 'Simili'.

Google | obama aborto | Cerca

Cerca: ☒ nel Web ☐ pagine in Italiano ☐ pagine provenienti da: Italia

Web [+ Mostra opzioni...](#) Ris

**[Obama firma ordine esecutivo su aborto in legge sanità ...](#)**  
25 mar 2010 ... **Obama** firma ordine esecutivo su **aborto** in legge sanità ... Le **opinioni** ed i commenti postati dagli utenti e le informazioni e dati in esso ...  
[www.diariodelweb.it/Articolo/Mondo/?d=20100325...](#) - [Copia cache](#)

**[Obama: dopo ok sanità, limiti su aborto - Top News - ANSA.it](#)**  
21 mar 2010 ... **Obama**: dopo ok sanità, limiti su **aborto**, Riaffermerà le limitazioni a uso di fondi federali per **aborto**, , Topnews, Ansa.  
[www.ansa.it/.../visualizza\\_new.html\\_1735732310.html](#) - [Copia cache](#)

**[L'aborto se vince Obama | club.quotidianonet.ilsole24ore.com](#)**  
Gaspari che, in due battute, ci dà l'idea del programma di **Obama** sui nascituri. La legge sull'**aborto** dei democratici Nel 2007 il candidato alla Casa Bianca ...  
[club.quotidianonet.ilsole24ore.com/.../l'aborto\\_se\\_vince\\_obama](#) - [Copia cache](#) - [Simili](#)

**[Staminali e aborto, svolta di Obama - LASTAMPA.it](#)**  
24 gen 2009 ... Il nuovo corso di Barack **Obama** imprime una decisa svolta alle ... una rimozione di tutti i limiti all'**aborto** decisi a livello federale e ...  
[www.lastampa.it/redazione/.../40332girata.asp](#) - [Copia cache](#) - [Simili](#)

**[Piccolino Valme: Con Obama aborto per tutti](#)**  
In merito alla legge che regola l'**aborto** negli Stati Uniti (Supreme Court Roe v. Wade), **Obama** ha più volte dichiarato, e lo ha scritto anche nelle pagine ...  
[piccolinovalme.blogspot.com/.../con-obama-aborto-per-tutti.html](#) - [Copia cache](#) - [Simili](#)

## 2. Opinioni di molti o.h.

- Qui i motori di ricerca fanno poco, al più identificano siti di opinioni

[Motorola V3 : Leggi le opinioni e compara i prezzi](#)

Motorola V3

20 gen 2008 ... **Motorola V3** a partire da EUR 66,00 (21.04.10) . Leggi 569 opinioni su **Motorola V3** e approfitta anche tu di questa incredibile offerta.

- Il metodo di ricerca dei motori web non è appropriato per l'analisi di opinioni
  - Il rank è basato sull'autorità di una pagina
  - In opinion mining spesso interessa una **statistica** sulle opinioni



## 2. Opinioni di molti o.h.(2)

- Quale tipo di ordinamento delle risposte è utile in opinion mining?
  - Aggregare opinioni positive e negative
  - Produrre un sommario (ma in che modo?)
  - Oppure, mostrare alcune opinioni più **RAPPRESENTATIVE**

# Una definizione del task di opinion mining

- E' innanzitutto un task di **classificazione**
- **Tipi di analisi:**
  - A livello di documento (globale)
  - A livello di singole frasi del documento
  - Al livello si singole features di oggetti (lo sterzo delle Ford)
  - A livello di un gruppo di documenti (sommario)
  - Paragoni, a livello di frasi o di features

## Una definizione del task di opinion mining (2)

- Componenti:
  - **Opinion holder**: chi esprime l'opinione
  - **Oggetto**: su cosa si esprime l'opinione
  - **Opinione**: un punto di vista, un apprezzamento, un'attitudine..

# Rappresentare gli oggetti e le opinioni

- Un oggetto può essere visto come appartenente ad una **gerarchia**, oppure un **cluster**, di oggetti
  - Turismo->Marocco->Riad->Riad Kniza
- Un oggetto può essere descritto come un insieme di attributi (**features**)
  - *Riad : servizi, personale, pulizia, luogo..*

## Rappresentare gli oggetti e le opinioni (2)

- Un oggetto  $O$  viene rappresentato mediante un insieme finito di features
$$F = \{f_1, f_2, \dots, f_n\}$$
- Ogni  $f_i$  può essere rappresentata da una o più keywords (sinonimi o categorie), es:  
“*luogo, posizione*”, oppure: “*servizi, sala conferenze, piscina, facilities, attrezzature, sala giochi..*”. Esiste dunque un set di termini per ogni feature
$$W = \{W_1, W_2, \dots, W_n\}$$
- Ogni oggetto o feature può essere descritto mediante una opinione  $P$
- Una opinione può anche essa essere rappresentata mediante un set di keywords (*opinion words*, **ow**)



# Tipi di analisi: Analisi a livello di documento e oggetto

Abbiamo scelto Riad Kniza sulla base di recensioni precedenti su Trip Advisor.

Abbiamo trascorso tre giorni e l'abbiamo trovata quasi **perfetto**. Il riad è **meravigliosamente** presentato ed è un rifugio completa di **tranquillità**. Il design è **impeccabile** e tutto è **ben organizzato**.

La nostra camera era - attenutata ma sentivo **eccellente** il letto era forse un po' **stretto** e la scelta di canali TV alquanto **limitata**.

Abbiamo mangiato solo la colazione e pranzo, ma abbiamo trovato cibo di **prima classe**. I pasti possono essere consumata sul tetto sotto o terrazza. Il servizio era **entusiasta** e **impeccabile**.

La posizione è **buona**, il riad è vicino al muro della medina e quindi i taxi sono solo un paio di minuti. È però a 10 minuti a piedi dal cuore della città.

Assegnando un peso alle espressioni di apprezzamento (**rosse**) e a quelle di critica (**blu**), si genera una valutazione complessiva (binaria o graduata) all'oggetto

## Analisi a livello di frase e

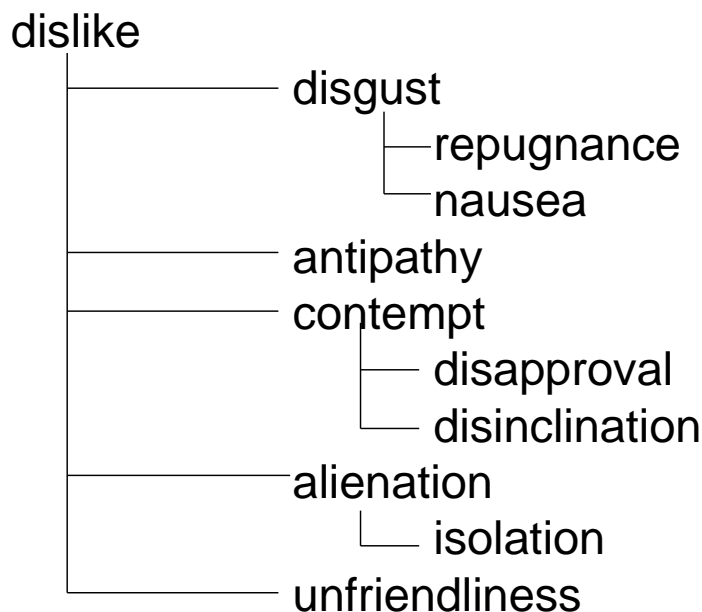
Abbiamo scelto **Riad Kniza** sulla base di recensioni precedenti su Trip Advisor. Abbiamo trascorso tre giorni e l'abbiamo trovato **quasi perfetto**. Il riad è meravigliosamente presentato ed è un rifugio completa di **tranquillità**. Il design è **impeccabile** e tutto è **ben organizzato**. La nostra **camera** era - attenutata ma sentivo eccellente i letto era forse un po' **strette** e la scelta di canali TV alquanto limitata. Abbiamo mangiato solo la colazione e pranzo, ma abbiamo trovato **cibo** di **prima classe**. I pasti possono essere consumata sul tetto sotto o terrazza. Il **servizio** era **entusiasta** e **impeccabile**. La **posizione** è **buona**, il riad è vicino al muro della medina e quindi i taxi sono solo un paio di minuti. È però a 10 minuti a piedi dal cuore della città.

Riad Kniza

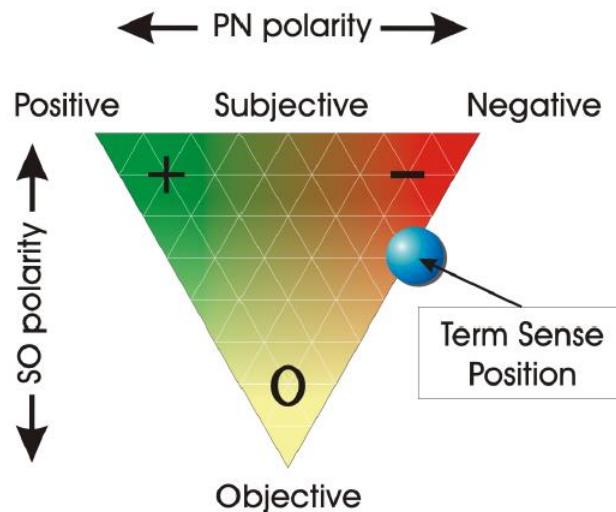
# Analisi a livello di features e frasi: quali compiti

1. Identificare ed estrarre gli attributi (**features**) sui quali l'opinion holder commenta (es. “**design**” vrs “**struttura**”, “**cibo**” vrs, “**colazione**”..): quali termini si associano quali featurers
2. Identificare le *opinion words* e associarle agli attributi (sentiment analysis) (es. **design** - >**impeccabile**)
3. Per determinare la connotazione di un'opinione (ad es + o -) occorre raggruppare termini in **gerarchie** o **classi di sinonimia**, poichè tanto le features **F** che le opinioni sono espresse con grande variabilità di termini **W**. Inoltre, spesso i termini usati per un'opinione dipendono dal tipo di feature (una **camera** è **bella**, il **cibo** è **buono**).
4. Produrre un sommario di opinioni (classificare)

## Risorse: WordNet Affect Taxonomy



# Risorse: SentiWordNet



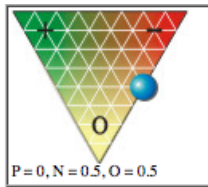
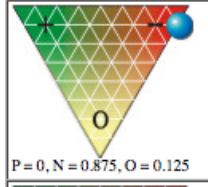
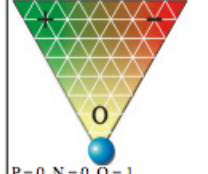
## SentiWordNet (complementare rispetto a WN Affect)

### Noun

3 senses found.

<p><math>P = 0.875, N = 0, O = 0.125</math></p>	<p><a href="#">good(2)</a> <a href="#">goodness(2)</a>  <i>moral excellence or admirableness; "there</i> </p>
<p><math>P = 0.5, N = 0, O = 0.5</math></p>	<p><a href="#">good(1)</a>  <i>benefit; "for your own good"; "what's the</i> </p>
<p><math>P = 0.75, N = 0, O = 0.25</math></p>	<p><a href="#">goodness(1)</a> <a href="#">good(3)</a>  <i>that which is good or valuable or useful; "self-realization"</i> </p>

- ..non completamente affidabile, perché 1) acquisito automaticamente 2)conserva l'ambiguità di WordNet

 <p>P = 0, N = 0.5, O = 0.5</p>	<p><a href="#">small(4)</a> <a href="#">little(4)</a>  <i>not fully grown; "what a big little boy you are"; "small children"</i></p>
 <p>P = 0, N = 0.875, O = 0.125</p>	<p><a href="#">low(7)</a> <a href="#">lowly(1)</a> <a href="#">modest(5)</a> <a href="#">small(3)</a> <a href="#">humble(1)</a>  <i>low or inferior in station or quality; "a humble cottage"; "a lowly parish priest"; "a</i></p>
 <p>P = 0, N = 0, O = 1</p>	<p><a href="#">small(10)</a> <a href="#">minuscule(2)</a> <a href="#">little(8)</a>  <i>lowercase; "little a"; "small a"; "e.e.cummings's poetry is written all in minuscule l</i></p>

## Riassunto

- Perché trovare opinioni è diverso dal trovare fatti
- Quali sono le entità coinvolte: oggetti, features, opinion words
- Quali sono i task: associare ow a features, classificare, sommarizzare, paragonare
- Modalità di analisi: document level, sentence level, feature level
- Next: survey dei metodi

# Survey di alcuni metodi usati in letteratura: **document level sentiment analysis**

## Document level sentiment analysis

- Turney ACL 2002: **Unsupervised review classification**
- Dati: reviews da *epinions.com* su automobili, banche, film, turismo..
- Tre passi:
  - **Step 1** analisi del testo (part-of-speech tagging, estrazione di coppie consecutive di parole con tags specifici (es Agg Nome), es “conveniently located” (l’idea è che si tratti di possibili features )

# Turney 2002 (cont'd)

- **Step 2:** assegnare una “semantic orientation” **SO** ad ogni coppia (*phrase*)
- Pointwise mutual information (per misurare la correlazione fra termini della coppia):

$$PMI(word_1, word_2) = \log_2 \left( \frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)} \right)$$

- SO(*phrase*) calcolata usando Altavista:  
 $SO(phrase) = PMI(phrase, \text{“excellent”})$   
-  $PMI(phrase, \text{“poor”})$  (i due aggettivi sono utilizzati per opinioni turistiche, diversi agg. per diversi dominii )

## Esempio

"conveniently located" excellent

Circa 9.090.000 risultati (0,13 secondi)

"conveniently located" poor

Circa 453.000 risultati (0,07 secondi)

Questo risultato suggerisce una connotazione **positiva** per “conveniently located”

# Turney 2002 (cont)

- **Step 3:** calcola il **valore medio** della SO di tutte le coppie estratte dal documento. Classifica l'opinione come “raccomandazione” se SO è positivo, come “non raccomandabile” altrimenti.
- Accuracy:
  - automobili - 84%
  - banche- 80%
  - film 65.83
  - destinazioni di viaggi - 70.53%

## Metodi supervisionati per document-level sentiment analysis

- Applicano tecniche di *machine learning*
- Partono da “datasets” di opinioni già classificate
- In fase di addestramento, il sistema impara la “polarità” positiva, negativa o neutrale delle parole sulla base di esempi
- In fase di classificazione, assegna una polarità ad un documento (vettore di features) sulla base della polarità delle parole in esso contenute
- Migliori prestazioni: SVM (83%)

# Sentence level sentiment analysis

- Step 1: Il primo passo consiste nel classificare le frasi come *soggettive* (cioè esprimenti un giudizio) o *oggettive*
- La maggior parte dei sistemi usa metodi di machine learning (es Bayesian classifier in Wiebe, ACL 1999)
- Alcuni metodi classificano patterns (sequenze specifiche di termini e o POS) anziché frasi ,  
es “<Noun> *was satisfied*”

## Sentence level sentiment analysis (2)

- Step 2: le frasi o patterns classificati come soggettivi vengono ulteriormente classificati come positivi, negativi, o neutri (ancora con metodi di machine learning, o metodi simili a Turney 2002)



# Feature-based opinion mining

- L'analisi di documenti o frasi assegna una polarità complessiva ad un **oggetto** o **servizio**, ma non identifica COSA un utente ha apprezzato o disprezzato
- Spesso l'opinione è articolata: “*servizio eccellente ma stanze troppo piccole*”
- L'analisi feature-based richiede di catturare in maniera più precisa i termini di “sentimento”

## Feature-based opinion mining(2)

- “Opinion words” o “opinion phrases”
    - **Positive**: beautiful, wonderful, good, amazing,
    - **Negative**: bad, poor, terrible, cost someone an arm and a leg (idiomatica)
    - **Context dependent**: “The battery life is *long*” (+) and “It takes a *long* time to focus” (-).
- Ci sono tre metodi per compilare questa lista di ow:
- A mano
  - Basandosi su corpora (ad esempio le congiunzioni di aggettivi sono spesso connotate emozionalmente)
  - Basandosi su dizionari (wordnet emotions, sentiwordnet)

# Feature-based opinion mining(3)

- L'analisi feature-based a livello di frasi è molto più raffinata, e presenta maggiore variabilità (parole diverse che identificano una feature, parole diverse che esprimono opinioni su features)
- Task1: identificare e estrarre le features dei vari oggetti
- Task 2: associare alle features opinion words
- Task 3: **creare liste di sinonimi** (per le features, oltre che per le opinion words)

Compito più o meno complesso a seconda del formato della review

## Format 1

My SLR is on the shelf

by camerafun4, Aug 09 '04

**Pros:** Great photos, easy to use, very small  
**Cons:** Battery usage; included memory is stingy.

I had never used a digital camera prior to purchasing have always used a SLR ... [Read the full review](#)

## Format 3

GREAT Camera., Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The **pictures** coming out of this camera are amazing. The **'auto'** feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out.

## Format 2

User  
rating  
Perfect  
10  
out of 10

"It is a great digital still camera for this century"

September 1, 2004

### Pros:

It's small in size, and the rotatable lens is great. It's very easy to use, and has fast response from the shutter. The LCD has increased from 1.5 in to 1.8, which gives bigger view. It has lots of modes to choose from in order to take better pictures.

### Cons:

It almost has no cons, it would be better if the LCD is bigger and it's going to be best if the model is designed to a smaller size.

# Feature-based

**GREAT Camera.**, Jun 3, 2004

Reviewer: **jprice174** from Atlanta, Ga.

I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.

The **pictures** coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. ...

- Classifica le varie frasi che parlano di una feature (Hu and Liu,

**Feature1: picture**

**Positive: 12**

- The **pictures** coming out of this camera are amazing.
- Overall this is a good camera with a really good **picture** clarity.

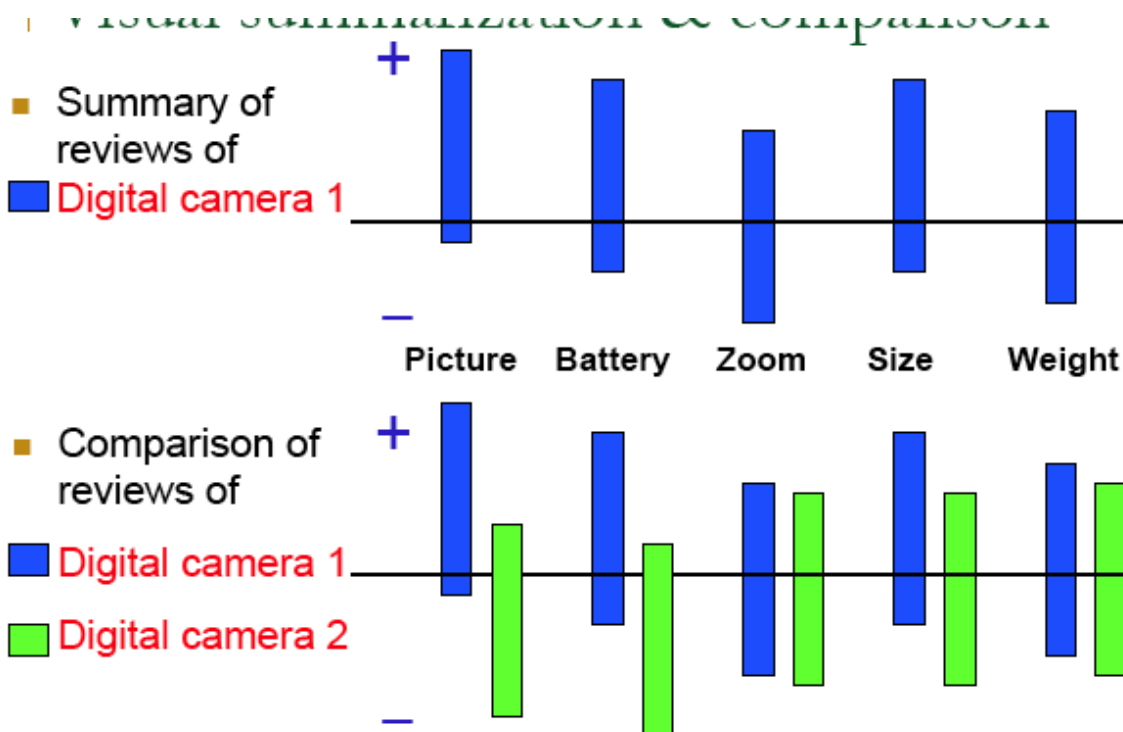
...

**Negative: 2**

- The **pictures** come out hazy if your hands shake even for a moment during the entire process of taking a picture.
- Focusing on a display rack about 20 feet away in a brightly lit room during day time, **pictures** produced by this camera were blurry and in a shade of orange.

**Feature2: battery life**

## Generare paragoni e sommari



# Feature-based analysis: problemi

- Identificare le features
- Identificare sinonimi di features
- Associare opinioni alle features

## Identificare le features: problemi

- Identificare le features (manualmente o mediante dizionari o con metodi statistici) è un task domain dependent, ogni dominio ha oggetti diversi e features diverse.
- Il modo con cui una *feature* è menzionata è variabile: “It is small enough to fit easily in a coat pocket or purse.” (la feature implicita è size!!)

## Identificare le features (2)

- **Metodi basati sulla frequenza:** alcune parole che rappresentano feature rilevanti tendono a apparire frequentemente (es: clean/cleaness per un albergo, size per una fotocamera, ecc.)
- Una volta identificate le keyword più rilevanti, possono essere estese con **sinonimi**, usando dizionari

## Identificare le features (3)

- Metodi basati sulle relazioni *part-of*: (Etzioni and Popescu, 2005) si estraggono dal web patterns che indicano relazioni part-whole rispetto ad un oggetto (es “*scanner comes with*” “*of scanner*” “*scanner has*”)  
*scanner has* an internal high frequency **power supply**  
*scanner comes with* no **cables** attached  
**effects of scanner**

# Identificare le features (4)

- Come trovare le features meno frequenti?
- Si analizzano espressioni di opinioni su features frequenti

“The pictures are absolutely amazing.”

- Si cercano espressioni uguali sostituendo alla feature un wildcard

“\* are absolutely amazing.”

- Si estraggono in tal modo features meno frequenti:

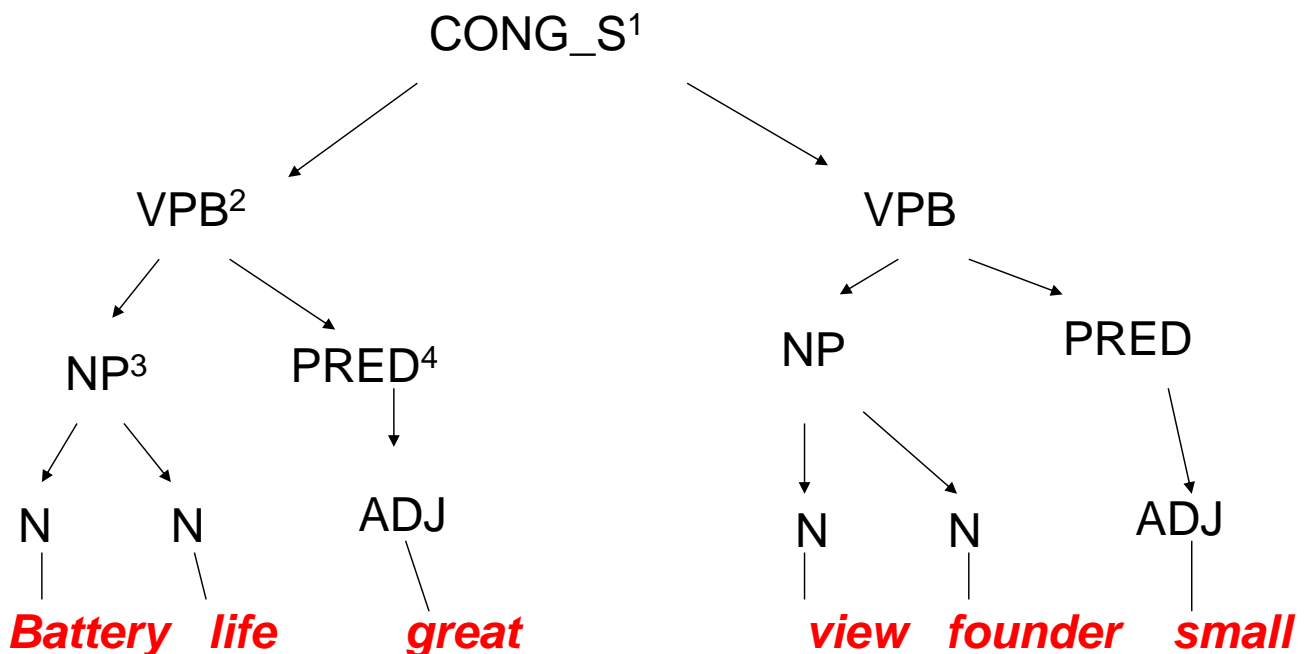
“The software that comes with it is amazing.”

## Identificare sinonimi di features

- Wordnet
- Lin similarity (modulo Perl scaricabile da: <http://kobesearch.cpan.org/htdocs/WordNet-Similarity/WordNet/Similarity/lin.html> )
- Molti algoritmi in letteratura. Guardate anche <http://webdocs.cs.ualberta.ca/~lindek/demos/>

# Associare opinioni alle features

- Per ogni feature, si identificano i termini esprimenti opinioni
- In genere a livello di frase, ma una frase può descrivere più features: The **battery life** and **picture quality** are **great** (+), but the **view founder** is **small** (-).
- Metodi più raffinati effettuano un parsing della frase per identificare quali attributi si riferiscono a quali features (un'analisi sintattica può associare il modificatore "great" alla prima frase della congiunzione, e "small" alla seconda).
- Alcuni *opinion terms* hanno un orientamento che dipende dal contesto (es **small** "+" per una macchina fotografica, "-" per una stanza di albergo)



1: conjunctive sentence (S but S)    4: predication

2: verb phrase *to be*

3: nopun phrase

# Determinare l'orientamento di opinioni su features: metodi

- Hu and Liu, KDD-04, Ding and Liu 2007
- Step 1: dividi la frase in segmenti  $s_i$ , separandola sulla base di congiunzioni (and, but, or..)
- Step 2: sia  $f_i$  la feature contenuta in  $s_i$ , e  $w_1, \dots, w_k$  le ow (*opinion words*) in  $s_i$



## Paragone fra features

Es: “*car X is not as good as car Y*”

- “comparative sentence mining”: implica due compiti:
  - Identificare espressioni comparative (“as **ow** as”, “more **ow** than”)
  - Determinare le relazioni di comparazione (“=“ “>” “<”).



## Paragone fra features (2)

- Le frasi comparative usano morfemi come : *more<most less/least, -er/-est*
- Ma le cose sono più complicate (*"In market capital, Intel is **way ahead** of Amd"*)

## Paragone fra features (3)

- Graduabili
  - Non uguaglianza (>,<): parole-chiave come *better, ahead, beats*.
  - Uguaglianza (=): *equal to, same as, both, all*
  - Superlativi: (er/est) *best, most. Better than all..*

## Paragone fra features (4)

- Esempi
- Ex1: “*car X has better controls than car Y*”  
(**relationWord** = better, features = controls, **entityS1** = car X, **entityS2** = car Y, **type** = non-equal-gradable)
- Ex2: “*car X and car Y have equal mileage*”  
(**relationWord** = equal, features = mileage, **entityS1** = car X, **entityS2** = car Y, **type** = equative)
- Ex3: “*Car X is cheaper than both car Y and car Z*” (**relationWord** = cheaper, features = **null**, **entityS1** = car X, **entityS2** = {car Y, car Z}, **type** = non-equal-gradable )
- Ex4: “*company X produces a variety of cars, but still best cars come from company Y*”( **relationWord** = best, **features** = cars, **entityS1** = company Y, **entityS2** = **null**, **type** = superlative)

## Paragone fra features (5)

(Jinal and Liu, SIGIR-06 e AAAI-06)

- Lista di 83 *keywords* usate in frasi comparative
- Step1: estrai frasi che contengono almeno una keyword (recall 98%, p 32%)
- Step 2: Naive Bayes o SVM classifier per classificare nelle 3 classi di “graduabilità”
- Step 3: estrazione delle relazioni

# Conclusioni

- Opinion Search: un'area nuova, di grande interesse applicativo, ancora molti problemi aperti
- Modello generale:
  - ENTITA' COINVOLTE: oggetti, features, opinion words
  - TASKS: analisi a livello di documento, a livello di frase, a livello di features, analisi comparative