What customers talk about?
Revealing hidden knowledge in large collections of reviews.
Understanding customers

• how customers perceive and evaluate products and services is an important element in improving business processes and increasing competitive advantage

• the question is how to get this knowledge
How to elicit opinions from customers?

- traditional methods like surveys where customers answer to a set of *predefined* questions
- the answers are evaluated, typically statistically
- problems that might occur
  - responses might be influenced by method bias when the respondents cannot provide accurate responses and/or when they are unwilling to provide accurate responses
  - the order of studied issues might play an important role
  - a precise specification of the properties of the interviewed subjects in order to ensure some representativeness
  - a risk of some important aspect omission
  - the number of processed responses is usually relatively low (a few hundred)
How to elicit opinions from customers?

• with the growth of volumes of electronic data, especially thanks to massive use of various on-line channels and platforms (social networks, discussion boards, on-line review sites), huge collections of documents containing customers’ opinions are available
Customer Reviews

★★★★★ 48
4.1 out of 5 stars

<table>
<thead>
<tr>
<th>Rating</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>77%</td>
</tr>
<tr>
<td>4 star</td>
<td>6%</td>
</tr>
<tr>
<td>3 star</td>
<td>2%</td>
</tr>
<tr>
<td>2 star</td>
<td>6%</td>
</tr>
<tr>
<td>1 star</td>
<td>9%</td>
</tr>
</tbody>
</table>

★★★★★ Great watch, great deal
By Jefferson Overlin on August 24, 2015
Verified Purchase

Great watch, great deal. Not a big fan of expensive stuff, but as a huge watch fan and space nerd I had to buy this. Well worth it. Very accurate watch and beautiful, simple design. If I could buy this watch 1000 times, I would!

4 of 4 people found the following review helpful

★★★★☆ Love this watch
By PAUL GLUB on January 6, 2015
Verified Purchase

Love this watch! However my biggest complaint is the Hesalite crystal scratches very easily. I mean you bump it wrong and you will see the scratch. Only wish it was more scratch resistant. Is keeping excellent time. This is something I will probably keep for many years and perhaps pass down to my son.
Customer Reviews

**4.1 out of 5 stars**

<table>
<thead>
<tr>
<th>Star Rating</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star</td>
<td>77%</td>
</tr>
<tr>
<td>4 star</td>
<td>6%</td>
</tr>
<tr>
<td>3 star</td>
<td>2%</td>
</tr>
<tr>
<td>2 star</td>
<td>6%</td>
</tr>
<tr>
<td>1 star</td>
<td>9%</td>
</tr>
</tbody>
</table>

---

Great watch, great deal

By Jefferson Overlin on August 24, 2015

Verified Purchase

Great watch, great deal. Not a big fan of expensive stuff, but as a huge watch fan and space nerd I had to buy this. Well worth it. Very accurate watch and beautiful, simple design. If I could buy this watch 1000 times, I would!

---

Love this watch

By PAUL GLUB on January 6, 2015

Verified Purchase

Love this watch! However my biggest complaint is the Hesalite crystal scratches very easily. I mean you bump it wrong and you will see the scratch. Only wish it was more scratch resistant. Is keeping excellent time. This is something I will probably keep for many years and perhaps pass down to my son.
product aspects and their perception can be obtained

- **accuracy**
  - (even expressed in different words)
- **design**
- **crystal durability**

**Customer Reviews**

- **Great watch, great deal**
  - By Jefferson Overlin on August 24, 2015
  - Verified Purchase
  
  Great watch, great deal. Not a big fan of expensive stuff, but as a huge watch fan and space nerd I had to buy this. Well worth it. Very accurate watch and beautiful, simple design. If I could buy this watch 1000 times, I would!

- **Love this watch**
  - By Paul GLUB on January 6, 2015
  - Verified Purchase
  
  Love this watch! However my biggest complaint is the Hesalite crystal scratches very easily. I mean you bump it wrong and you will see the scratch. Only wish it was more scratch resistant. Is keeping excellent time. This is something I will probably keep for many years and perhaps pass down to my son.
Janet
United Kingdom
1 review

“Friendly, clean, and well worth a visit”

Leisure trip  Solo traveler  Double Room  Stayed 6 nights  Submitted from a mobile device

+ Good position, no problem with parking, nice friendly staff especially I do not speak French, and comfortable

Helpful

Yumi
Austria
1 review

“Top value for money: recommended!”

Leisure trip  Couple  Double Room  Stayed 1 night  Submitted from a mobile device

- Elevator was missing

+ After enjoying an Ibis budget hotel in Marseille we were more than surprised what you can get for the money: very clean, new, affordable, truly recommendable.

Helpful
How to elicit opinions from customers?

• with the growth of volumes of electronic data, especially thanks to massive use of various on-line channels and platforms (social networks, discussion boards, on-line review sites), huge collections of documents containing customers’ opinions are available

• manual analysis of the data by linguistic and domain experts is demanding and often complicated
Stuhrling Original Men's 127A.3315C2 Special Reserve Emperor's Grandeur Automatic Dual Time Skeleton Dial Blue Leather Strap Watch

* * * * * 147 customer reviews

List Price: $995.00
Price: $177.44 & FREE Shipping. FREE Returns.
Details
You Save: $817.56 (82%)

In Stock.
Want it tomorrow, Jan. 23? Order within 5 hrs 26 mins and choose Saturday Delivery at checkout. Details
reviews in many languages
very many reviews
reviews in many languages
= additional problem
very many reviews

reviews in many languages = additional problem or opportunity
How to elicit opinions from customers?

• with the growth of volumes of electronic data, especially thanks to massive use of various on-line channels and platforms (social networks, discussion boards, on-line review sites), huge collections of documents containing customers’ opinions are available

• manual analysis of the data by linguistic and domain experts is demanding and often complicated

• processing the data within a reasonable time and budget is not feasible
The problem

- existence of a large collection of short Internet textual documents (freely written in various natural languages) related to a certain service (accommodation)
- texts labeled as positive and negative
- manual processing not feasible
- the goal – to reveal what causes positive or negative impressions automatically
What is important?

• the reasons of a sentiment is expressed by words contained in the reviews – the question is: “Which of them are important?”
• it is not possible to simply look only at word frequencies
• it is a feature selection problem (supervised – we have the labels)
Attribute selection

• filter methods
  – don’t consider any classifier, rank attributes
  – e.g., $\chi^2$, Information gain
  – don’t consider relationships between attributes
  – give no information about usefulness of the attributes
Attribute selection

• decision tree learner
  – induces a decision tree that might be used to classify documents
  – the c4.5 algorithm successively selects attributes that best split the data into more and more homogeneous groups (expressed by entropy)
  – attributes are included in the nodes of the DT
  – after pruning the DT only the most important ones remain (e.g., 200-300 attributes from 25,000 are identified as relevant)
  – accuracy around 90%
### Important attributes revealed

<table>
<thead>
<tr>
<th>English</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>100</td>
</tr>
<tr>
<td>friendly</td>
<td>80</td>
</tr>
<tr>
<td>not</td>
<td>77</td>
</tr>
<tr>
<td>excellent</td>
<td>73</td>
</tr>
<tr>
<td>helpful</td>
<td>67</td>
</tr>
<tr>
<td>spacious</td>
<td>62</td>
</tr>
<tr>
<td>friendliness</td>
<td>59</td>
</tr>
<tr>
<td>beautiful</td>
<td>57</td>
</tr>
<tr>
<td>comfortable</td>
<td>55</td>
</tr>
<tr>
<td>nice</td>
<td>55</td>
</tr>
<tr>
<td>conveniently</td>
<td>54</td>
</tr>
<tr>
<td>convenient</td>
<td>53</td>
</tr>
<tr>
<td>fantastic</td>
<td>53</td>
</tr>
<tr>
<td>good</td>
<td>52</td>
</tr>
<tr>
<td>proximity</td>
<td>52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>German</th>
<th>%</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>lage</td>
<td>100</td>
<td>location</td>
</tr>
<tr>
<td>gutes</td>
<td>78</td>
<td>good</td>
</tr>
<tr>
<td>nicht</td>
<td>76</td>
<td>no</td>
</tr>
<tr>
<td>freundliche</td>
<td>74</td>
<td>friendly</td>
</tr>
<tr>
<td>freundliches</td>
<td>73</td>
<td>friendly</td>
</tr>
<tr>
<td>freundlich</td>
<td>70</td>
<td>friendly</td>
</tr>
<tr>
<td>nettes</td>
<td>65</td>
<td>nice/kind</td>
</tr>
<tr>
<td>nette</td>
<td>64</td>
<td>nice/kind</td>
</tr>
<tr>
<td>zentral</td>
<td>62</td>
<td>central</td>
</tr>
<tr>
<td>gute</td>
<td>61</td>
<td>good</td>
</tr>
<tr>
<td>schönes</td>
<td>61</td>
<td>nice</td>
</tr>
<tr>
<td>freundlichkeit</td>
<td>59</td>
<td>friendliness</td>
</tr>
<tr>
<td>zuvorkommend</td>
<td>56</td>
<td>helpful</td>
</tr>
<tr>
<td>tolles</td>
<td>56</td>
<td>great</td>
</tr>
<tr>
<td>ruhig</td>
<td>55</td>
<td>quiet</td>
</tr>
</tbody>
</table>
Relation of attributes to classes

LOCATION > 0:
  ....POOR > 0:
  :   ....GOOD > 0: POS (13)
  :   :   GOOD <= 0:
  :   :     EXCELLENT > 0: POS (3)
  :   :     EXCELLENT <= 0:
  :   :       GREAT > 0: POS (3)
  :   :       GREAT <= 0:
  :       :     CLEAN <= 0: NEG (48/4)
  :       :       CLEAN > 0: POS (4/1)
  :   POOR <= 0:
  :     DIFFICULT > 0:
  :       ....GOOD > 0: POS (6)
  :       :   GOOD <= 0:
  :       :     HELPFUL <= 0: NEG (34/7)
  :       :       HELPFUL > 0: POS (5)
  :   ....HELPFUL > 0: POS (5)

analysis of DT branches reveals words significant for each class
Weighting the attributes

- ideal situation – points on the diagonal
- $w_3$ is better than $w_1$
- situation with $w_1$ and $w_2$ is not so clear
Weighting the attributes

- introducing a weight balancing the frequency when a word was used for classification and the correctness of the classification
- the weight determines the importance of a word in relation to a given category (positive or negative class) – higher numbers mean bigger relevancy.

\[ w_w = \frac{N_C}{N_N} \times \frac{\ln(\sqrt{N_C^2 + N_N^2})}{\ln(N_{\text{max}})} \]
<table>
<thead>
<tr>
<th>English</th>
<th>Positive</th>
<th>Negative</th>
<th>German</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Translation</td>
<td>Original</td>
<td>Translation</td>
<td>Original</td>
<td>Translation</td>
</tr>
<tr>
<td>location</td>
<td>-</td>
<td>no</td>
<td>-</td>
<td>lage</td>
<td>location</td>
</tr>
<tr>
<td>friendly</td>
<td>-</td>
<td>not</td>
<td>-</td>
<td>gutes</td>
<td>good</td>
</tr>
<tr>
<td>excellent</td>
<td>-</td>
<td>small</td>
<td>-</td>
<td>freundlich</td>
<td>friendly</td>
</tr>
<tr>
<td>helpful</td>
<td>-</td>
<td>bit</td>
<td>-</td>
<td>freundliche</td>
<td>friendly</td>
</tr>
<tr>
<td>comfortable</td>
<td>-</td>
<td>noise</td>
<td>-</td>
<td>freundliches</td>
<td>friendly</td>
</tr>
<tr>
<td>nice</td>
<td>-</td>
<td>poor</td>
<td>-</td>
<td>gute</td>
<td>good</td>
</tr>
<tr>
<td>spacious</td>
<td>-</td>
<td>noisy</td>
<td>-</td>
<td>freundlichkeit</td>
<td>friendliness</td>
</tr>
<tr>
<td>beautiful</td>
<td>-</td>
<td>could</td>
<td>-</td>
<td>nette</td>
<td>nice</td>
</tr>
<tr>
<td>good</td>
<td>-</td>
<td>would</td>
<td>-</td>
<td>nettes</td>
<td>nice</td>
</tr>
<tr>
<td>convenient</td>
<td>-</td>
<td>expensive</td>
<td>-</td>
<td>zentral</td>
<td>central</td>
</tr>
<tr>
<td>staff</td>
<td>-</td>
<td>t</td>
<td>-</td>
<td>schönes</td>
<td>nice</td>
</tr>
<tr>
<td>wonderful</td>
<td>-</td>
<td>too</td>
<td>-</td>
<td>super</td>
<td>super</td>
</tr>
<tr>
<td>clean</td>
<td>-</td>
<td>should</td>
<td>-</td>
<td>schöne</td>
<td>nice</td>
</tr>
<tr>
<td>amazing</td>
<td>-</td>
<td>little</td>
<td>-</td>
<td>ruhig</td>
<td>quiet</td>
</tr>
<tr>
<td>close</td>
<td>-</td>
<td>need</td>
<td>-</td>
<td>hotel</td>
<td>hotel</td>
</tr>
<tr>
<td>friendliness</td>
<td>-</td>
<td>lack</td>
<td>-</td>
<td>zuvorkommend</td>
<td>helpful</td>
</tr>
<tr>
<td>helpfull</td>
<td>-</td>
<td>limited</td>
<td>-</td>
<td>netter</td>
<td>nice</td>
</tr>
<tr>
<td>proximity</td>
<td>-</td>
<td>be</td>
<td>-</td>
<td>sauberer</td>
<td>clean</td>
</tr>
<tr>
<td>enjoyed</td>
<td>-</td>
<td>cold</td>
<td>-</td>
<td>nett</td>
<td>nice</td>
</tr>
<tr>
<td>relaxing</td>
<td>-</td>
<td>difficult</td>
<td>-</td>
<td>angenehme</td>
<td>pleasant</td>
</tr>
</tbody>
</table>
Extracting significant expressions

- more informative than single words
- a significant expression = frequent combination of words in the proximity a significant word

The triplets found (potential significant expressions):
- helpfulness of the staff
- helpfulness of staff
- helpfulness of at
- helpfulness of check
- helpfulness of in
- helpfulness the staff
- helpfulness the at
- helpfulness the check
- helpfulness the in
- etc.
<table>
<thead>
<tr>
<th>Original</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>und das Zimmer war gross</td>
<td>and the room was big</td>
</tr>
<tr>
<td>Personal sehr freundlich und zuvorkommend</td>
<td>staff very friendly and helpful</td>
</tr>
<tr>
<td>man kann alles zu Fuss erreichen</td>
<td>one can reach everything by walk</td>
</tr>
<tr>
<td>das Zimmer und bad sehr sauber</td>
<td>the room and bathroom very clean</td>
</tr>
<tr>
<td>ein Paar Schritten vom Strand Weg</td>
<td>a few steps from beach path</td>
</tr>
<tr>
<td>Terrasse mit Blick auf das Meer</td>
<td>terrace with view of the sea</td>
</tr>
<tr>
<td>sehr gute Lage ansprechendes Design</td>
<td>very good location attractive</td>
</tr>
<tr>
<td></td>
<td>design</td>
</tr>
<tr>
<td>die Dusche war gross und schön</td>
<td>shower was big and nice</td>
</tr>
<tr>
<td>das Hotel is zentral gelegen</td>
<td>hotel is centrally located</td>
</tr>
<tr>
<td>das Frühstück war sehr gut</td>
<td>breakfast was very good</td>
</tr>
<tr>
<td>Original</td>
<td>Translation</td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>die Bar und das Restaurant geschlossen</td>
<td>the bar and the restaurant closed</td>
</tr>
<tr>
<td>Duschen und Waschbecken sind sehr klein</td>
<td>shower and basin are very small</td>
</tr>
<tr>
<td>die Klimaanlage nicht funktioniert</td>
<td>air-conditioning not working</td>
</tr>
<tr>
<td>Zimmer extrem klein für den Preis</td>
<td>extremely small room for the price</td>
</tr>
<tr>
<td>die Sterne Hotels nicht entsprechen</td>
<td>stars of the hotel don’t correspond</td>
</tr>
<tr>
<td>so das war nicht schön Zimmer</td>
<td>so it was not a nice room</td>
</tr>
<tr>
<td>ist sehr schwer finden das Hotel</td>
<td>is difficult to find the hotel</td>
</tr>
<tr>
<td>die Fenster waren nicht sauber</td>
<td>windows were not clean</td>
</tr>
<tr>
<td>Zimmer zur Strasse sind sehr laut</td>
<td>rooms to the street are noisy</td>
</tr>
<tr>
<td>für Leute mit Rückenleiden Betten nich</td>
<td>no beds for people with back problems</td>
</tr>
</tbody>
</table>
Mining n-grams

• working with sequences containing n words instead of single words
• keeping the original word order
• using a similar procedure as mentioned before (used rule generator instead of a decision tree, n-grams rated according to rule lift)
• including also “complementary” knowledge from the rules of induced decision tree
• e.g., \( IF \) very good breakfast = 0 AND owners very friendly = 0 AND room very good = 0 \( THEN \) class = negative means that very good breakfast is positive
<table>
<thead>
<tr>
<th>English positive</th>
<th>English negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>good value money</td>
<td>could hear everything</td>
</tr>
<tr>
<td>hotel very clean</td>
<td>bed very hard</td>
</tr>
<tr>
<td>breakfast very good</td>
<td>no hot water</td>
</tr>
<tr>
<td>staff very nice</td>
<td>not very clean</td>
</tr>
<tr>
<td>very good value</td>
<td>very noisy night</td>
</tr>
<tr>
<td>close train station</td>
<td>no free internet</td>
</tr>
<tr>
<td>very good location</td>
<td>bit too small</td>
</tr>
<tr>
<td>within walking distance</td>
<td>not very good</td>
</tr>
<tr>
<td>very nice hotel</td>
<td>not good value</td>
</tr>
<tr>
<td>great value money</td>
<td>(not) very good breakfast</td>
</tr>
<tr>
<td>German positive / translation</td>
<td>German negative / translation</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>freundliches hilfsbereites personal / friendly helpful staff</td>
<td>zimmer extrem kleine / extremely small room</td>
</tr>
<tr>
<td>(not) stark befahrenen strasse / (not) bussy street</td>
<td>fenster nicht öffnen / unopenable window</td>
</tr>
<tr>
<td>(not) keine klimaanlage zimmer / air-conditioned rooms</td>
<td>rezeption nicht besetzt / no one in reception</td>
</tr>
<tr>
<td>personal beim frühstück / staff at breakfast</td>
<td>frühstück nicht gut / breakfast not good</td>
</tr>
<tr>
<td>(not) zimmer bad klein / (not) small bathroom</td>
<td>kein warmes wasser / no hot watter</td>
</tr>
<tr>
<td>vier sterne hotel / four-star hotel</td>
<td>hotel schwer finden / difficult to find hotel</td>
</tr>
<tr>
<td>(not) nicht wirklich sauber / really clean</td>
<td>zimmer strasse laut / room at noisy street</td>
</tr>
<tr>
<td>(not) frühstück nicht gut / good breakfast</td>
<td>zimmer klein bad / with small bathroom</td>
</tr>
<tr>
<td>(not) keine minibar zimmer / rooms with minibars</td>
<td>zimmer recht klein / room quite small</td>
</tr>
<tr>
<td>(not) kein kühlschrank zimmer / rooms with fridges</td>
<td>kein wi-fi lan / no wi-fi network</td>
</tr>
</tbody>
</table>
Unlabeled data

• What if class labels are not available?

⭐⭐⭐⭐⭐ Proved to be somewhat of a disappointment
By Teapot Tales on September 5, 2011
Although we all know that the Portuguese time pieces are by all means more valuable and of much higher quality than the Swiss time pieces, but all in all, this watch has only one truly redeeming quality - its fine leather band. Other than that, I became terribly disappointed about this watch..... I soon found out that my IWC time piece has a crystal that sticks too far out and gets scratched all the time as I enter or leave any one of my Maybachs, Koenigsegs and Aston Martins. ...

• What if we don’t want to analyze the reasons of a sentiment but rather other product/service aspects?
Mining relevant information from unlabeled data

- we have a lot of reviews discussing a product/service, we want to know what are the important aspects as perceived by customers and to understand what characterizes these aspects from the semantic point of view
- at the beginning, the included aspects (topics) are unknown → they must be separated
- than, representative characteristics of the separated groups must be extracted
- a human expert than decides about usefulness of the gained knowledge
The data

• the procedure was tested with real world data
• tens of thousands of customer reviews from
  – Booking.com
  – Amazon.com
Nicholas

7.9 “Comfortable stay”

- Leisure trip  - Couple  - Double Room  - Stayed 1 night

- Slightly out of town having been called Frejus Centre
- A welcoming manager and friendly staff, room small but comfortable. Good breakfast

Helpful

---

Janet

10 “Friendly, clean, and well worth a visit”

- Leisure trip  - Solo traveler  - Double Room  - Stayed 6 nights  - Submitted from a mobile device

+ Good position, no problem with parking, nice friendly staff especially I do not speak French, and comfortable

Helpful

---

Yumi

9.2 “Top value for money: recommended!”

- Leisure trip  - Couple  - Double Room  - Stayed 1 night  - Submitted from a mobile device

- Elevator was missing
- After enjoying an Ibis budget hotel in Marseille we were more than surprised what you can get for the money: very clean, new, affordable, truly recommendable.

Helpful
I was surprised at this phone. I was worried cuz I hadn’t ever heard of the brand. But I’ve had it for about 3 weeks now and I love it. The battery life is great considering how much I use it! It’s got an awesome camera, front and back. It’s lightweight. The screen is crisp and clear, colors are vibrant. Only a few complaints. There’s no app launcher, but it’s an easy fix. You can download one from the playstore. Also, you can’t change the background on the lock screen, but that has nothing to do with the phone itself. Over all, I’m very impressed and would recommend for anyone looking for a decent priced smart phone.

Won’t turn on, won’t charge, wouldn’t connect to network when it WAS working.

Purchased the phone a few days ago and tried it on the T-Mobile and Ting networks (same network). It will not hold on to a signal. No matter wherever I am it constantly fluctuates from perfect signal to no signal every few seconds. It won’t stay connected to a mobile data network for more than a few seconds at a time. Other phones on the same network sitting next to this one are not having the same problem, and it isn’t local… I traveled out of town and had the same problem. I’ll try a new one, but if it has the same problem I won’t be buying another Blu product.

I exchanged the phone and the new one did the exact same thing. It’s clearly a problem with the phones and not the carrier. A phone that won’t stay connected to a mobile network is useless.

Also, the product listing promised an upgrade to Android 5.0…but you can’t update it to Android 5.0, even Blu says you can’t.

EDIT: Now, after a week of owning this phone, it won’t turn on or charge. I can’t even send it back because there is no way to do a factory reset on it if it won’t turn on.

I won’t be buying another Blu product, ever.
The data

• preprocessing
  – case folding, keeping only regular words, removing rare terms and stop words (e.g., articles, prepositions, pronouns), stemming, splitting into sentences, creating n-grams, ...
  – converting to the structured vector space model
Documents clustering

• a way to separate groups of similar documents – an answer to unavailability of labeled data
• grouping data into subsets using only information contained in the data
• a problem – no labels, i.e., no one correct solution (number of clusters, assignment of documents to clusters)
Documents clustering

• flat vs. hierarchical clustering
  – flat clustering is suitable for large data sets

• clustering parameters
  – algorithm – k-means
  – criterion function – hybrid
  – similarity measure – cosine similarity
Documents clustering

• finding a correct number of clusters
  – given in advance
    • a known value
    • a manageable value
  – calculated
    • based on values derived from the data
The elbow method – adding another cluster doesn’t give much better modeling of the data set
Documents clustering

• evaluating clustering results
  – automatic evaluation
    • can give some information about homogeneity of the partitioned data
  – expert based evaluation
    • can evaluate usefulness
    • the expert needs some assistance
Revealing significant cluster representatives

• examining only some of the documents
  – e.g., average documents (near cluster centers)
• application of a statistical (e.g., the $\chi^2$ statistics) or machine learning method (e.g., a decision tree learner) to reveal typical characteristics (attributes) of the clusters
• from the extracted information an expert might make a conclusion
analysis of decision tree branches can reveal what words are relevant for every class
Examples of results

• we tested the procedure on two data sets
• we tried different approaches (algorithms, parameters, preprocessing, ...) to get acceptable results
• in all cases, we obtained groups of reviews sharing a common topic that could be revealed (characterized) with some assistance
Derived topics and representative reviews – cell phones (amazon.com)

• Mobile accessory (batteries)
  – After replacing my battery my phone no longer worked and I had to buy a new one so I would not recommend this.
  – This is a good battery. It works 90% like the original battery. If you need a replacement battery this will make a good one.

• Mobile accessory (chargers)
  – This is a perfect charger for my car. It works great, and the price is right. I recommend this product to all.
  – Works well, charges phone quickly, easy to use. Would recommend to others who need a charger. I would buy again.
## Derived topics and representative words – cell phones (amazon.com)

<table>
<thead>
<tr>
<th>Important attributes</th>
<th>Derived topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>batteri, charg, life, mah, evo, hour, origin, oem, stock, die</td>
<td>Mobile accessory (battery)</td>
</tr>
<tr>
<td>charger, cord, car, retract, plug, charg, work, wall, usb, transmitt</td>
<td>Mobile accessory (charger)</td>
</tr>
<tr>
<td>sound, ear, headset, hear, bluetooth, nois, comfort, music, pair, listen</td>
<td>Mobile accessory (headset, sound)</td>
</tr>
<tr>
<td>cabl, lg, usb, comput, transfer, data, tracfon, charg, pc, micro</td>
<td>Mobile accessory (cables, connectivity)</td>
</tr>
<tr>
<td>money, wast, worth, save, spend, junk, dont, buy, don, total</td>
<td>Customers discussing price</td>
</tr>
<tr>
<td>glare, film, anti, matt, finish, screen, mirror, protector, fingerprint, retina</td>
<td>Mobile accessory (screen film and protection)</td>
</tr>
<tr>
<td>color, pink, white, love, pictur, purpl, case, black, yellow, green</td>
<td>Customers discussing colors</td>
</tr>
</tbody>
</table>
Derived topics and representative words – hotel reviews (booking.com)

**Spanish**
- habitaciones (rooms)
- vistas (views)
- son (are)
- pequeñas (small)
- camas (beds)
- las (the)
- instalaciones (facilities)
- comodas (comfortable)
- como (as)

**Italian**
- cortesia (kindness)
- disponibilità (availability)
- tranquillità (tranquility)
- pulizia (cleanliness)
- accoglienza (welcome)
- gentilezza (kindness)
- personale (staff)
- del (of)
- cordialità (friendliness)
- silenzio (silence)
Derived topics and representative words – hotels reviews (booking.com)

**English**
- comfortable
- spacious
- clean
- bed
- modern
- quiet
- room
- large
- lovely
- well

**French**
- salle (room)
- peu (small)
- trop (too)
- manque (lack)
- absence (absence)
- odeur (smell odor)
- douche (shower)
- insonorisation (noise insulation)
- bruit (noise)
- chambre (room)
Topics derived from representative words – hotels reviews (booking.com)

- **CS**: general positives, breakfast, positives (no diacritic), staff, room negatives, staff/cleanliness, location, surroundings
- **DE**: general positives, location, breakfast, general positives, room positives, quality/price, atmosphere
- **EN**: value, hotel facilities, room positives, room negatives, staff, room facilities, location
- **ES**: rooms, environment, location (no diacritic), location (with diacritic), general negatives, rooms, quality/price, location
- **FR**: breakfast, facilities, environment, room negatives, location, price, quality, comfort
- **IT**: location, staff, facilities, rooms, convenience, room facilities, location, good quality, room positives
Conclusions

• the procedure could help in revealing topics (aspects, as perceived by customers) hidden in large collections of textual documents (customer reviews) related to a certain group of products or services

• the procedure did not require a domain knowledge, additional linguistic information, or lexicons
Conclusions

• strongly experimental work with no perfect output (even human experts don’t have to agree)

• using different algorithms, their parameters, including or excluding some preprocessing steps more or less changes the results (e.g., not excluding stop words revealed a group of reviews related to females)
Vielen Dank für Ihre Aufmerksamkeit.
Thank you for your attention.
Gracias por vuestra atención.
Спасибо за Ваше внимание.
Děkuji za pozornost.