Micropinion Generation: An Unsupervised Approach to Generating Ultra-Concise Summaries of Opinions

Kavita Ganesan, ChengXiang Zhai & Evelyne Viegas

Go to project page
Opinion Summarization Today...

Current methods: Focus on generating aspect based ratings for an entity

[Lu et al., 2009; Lerman et al., 2009;..]

### Customer Reviews

<table>
<thead>
<tr>
<th>Average Customer Rating</th>
<th>(1,432 customer reviews)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 star:</td>
<td>(1,040)</td>
</tr>
<tr>
<td>4 star:</td>
<td>(227)</td>
</tr>
<tr>
<td>3 star:</td>
<td>(63)</td>
</tr>
<tr>
<td>2 star:</td>
<td>(25)</td>
</tr>
<tr>
<td>1 star:</td>
<td>(77)</td>
</tr>
</tbody>
</table>

### Opinion Summary for iPod Touch

- **Appearance**: ★★★★★ (1,213)
- **Ease of use**: ★★★★★ (1,212)
- **Portability**: ★★★★★ (1,202)
- **Sound quality**: ★★★★★ (1,196)

> See and rate all 1 attributes.

### Most Helpful Customer Review

3,677 of 3,770 people found the following review helpful.

**WARNING** for new iPod Touch.

By Hassan B. Bn Hadhram

This review is from: Apple iPod touch 8 GB (2nd Generation—With iPhone OS 3.1 Software Installed) [NEWEST MODEL] (Electronics)

Before I start let me just tell you "what's New" with the iPod touch Third generation:

- Faster Cpu/Double the ram/Better graphic (faster Boot time/faster loading is all what i did notice)
- Double the storage for the same old price
- Voice control (I'll explain it in a second)
- Latest firmware for free
Opinion Summary for iPod Touch

To know more: read many redundant sentences

Most Helpful Customer Reviews

3,677 of 3,770 people found the following review helpful:

⭐⭐⭐⭐⭐ WARNING for new 8GB 3G owners and iPod touch 3G Review, September 11, 2009

By Hassan B. Bn Hadhram - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: Apple iPod touch 8 GB (2nd Generation--with iPhone OS 3.1 Software Installed) [NEWEST MODEL] (Electronics)

Before i start let me just tell you "what's New" with the iPod touch Third generation:

- Faster CPU/Double the ram/Better graphic (faster Boot time/faster loading is all what i did notice)
- Double the storage for the same old price
- Voice control (I'll explain it in a second)
- Latest firmware for free
Opinion Summarization Today...

Opinion Summary for iPod Touch

Structured summaries useful, but insufficient!

To know more: read many redundant sentences

Customer Reviews

Average Customer Rating

- Appearance: 4 stars (1,213)
- Ease of use: 4 stars (1,212)
- Portability: 5 stars (1,202)
- Sound quality: 4 stars (1,196)

Most Helpful Customer Reviews

3,677 of 3,770 people found the following review helpful:

⭐️⭐️⭐️⭐️⭐️ WARNING for new 8GB 3G owners and ipod touch 3G Review, September 11, 2009
By Hassan B. Bn Hadhram - See all my reviews

Amazon
This review is from:IPod Touch 3G 8GB (Electronics)
Textual opinion summaries:
• big and clear screen
• battery lasts long without wifi
• great selection of apps

Textual summaries can provide more information...
Properties of useful textual summaries

- **Represent the major opinions**
  - Key complaints/praise in text
  - Important critical information

- **Readable/well formed**
  - Easily understood by readers

- **Compact/Concise**
  - Can be viewed on all screen sizes
    - e.g. phone, pda, tablets, dekstops
  - Maximize information conveyed
Goal of this summarization work

- Generate a set of non-redundant phrases:
  - Summarizing **key opinions** in text
  - Short (2-5 words)
  - Readable

**Micropinions**

**Micropinion summary for a restaurant:**

“Good service”
“Delicious soup dishes”

- Ultra-concise nature of phrases to allow flexible adjustment of summaries according to display constraints
How to generate such ultra-concise summaries?
1. Extractive summarization

- Has been widely studied
  [Radev et al. 2000; Erkan & Radev, 2004; Mihalcea & Tarau, 2004...]

- Fairly easy to implement

- **Problem**: Not suitable for concise summaries
  - **Bias** – with limit on summary size
    - selected sentence/phrase may have missed critical info
  - **Verbose** - may contain irrelevant information
    - not suitable for smaller devices
2. Pure abstractive summarization

- Understand the original text and “re-tell” story in a fewer words ➔ **Hard to achieve!**

- Some methods require **manual effort**
  [DeJong 1982; Radev & McKeown 1998; Finley & Harabagiu 2002]
  - Need to define **templates**
  - Later filled with info using IE techniques

- Some methods rely on **deep NL understanding**
  [Saggion & Lapalme 2002; Jing & McKeown 2000]
  - Domain dependent
  - Impractical – high computational costs
3. Keyphrase extraction approaches

- Goal is to extract **important phrases** from text
  - Traditionally used to characterize documents
  - Can be potentially used to select key opinion phrases
  - Closest to our goal!

- **Problem:**
  - Only **topic phrases** may be selected
    - E.g. battery life, screen size ➔ candidate phrases
    - Enough to characterize documents, but we need more info!
  - **Readability** aspect is not much of a concern
    - E.g. "Battery short life" == "Short battery life"
  - Most methods are **supervised** – need training data
    [Tomokiyo & Hurst 2003; Witten et al. 1999; Medelyan & Witten 2008]
We propose...

- Unsupervised, lightweight & general approach to generating ultra-concise summaries of opinions

- Idea is to use existing words in original text to compose meaningful summaries

- Emphasis on 3 aspects:
  - **Compactness**
    - summaries should use as few words as possible
  - **Representativeness**
    - summaries should reflect major opinions in text
  - **Readability**
    - summaries should be fairly well formed
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \sum_{i=1}^{k} S_{\text{rep}}(m_i) + S_{\text{read}}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{\text{rep}}(m_i) \geq \sigma_{rep} \]

\[ S_{\text{read}}(m_i) \geq \sigma_{read} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{\text{sim}} \forall i, j \in [1, k] \]
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \left\{ m_i \ldots m_k \right\} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{rep}(m_i) \geq \sigma_{rep} \]

\[ S_{read}(m_i) \geq \sigma_{read} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k] \]
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \{ m_i \ldots m_k \} \sum_{i=1}^{k} S_{\text{rep}}(m_i) + S_{\text{read}}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{\text{rep}}(m_i) \geq \sigma_{\text{rep}} \]

\[ S_{\text{read}}(m_i) \geq \sigma_{\text{read}} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{\text{sim}} \forall i, j \in [1, k] \]

Micropinion Summary, M

2.3 very clean rooms
2.1 friendly service
1.8 dirty lobby and pool
1.3 nice and polite staff
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \{m_i \ldots m_k\} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{rep}(m_i) \geq \sigma_{rep} \]

\[ S_{read}(m_i) \geq \sigma_{read} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k] \]
Optimization Framework to capture compactness, representativeness & readability

\[
M = \arg \max \{m_i \ldots m_k\} \sum_{i=1}^{k} S_{\text{rep}}(m_i) + S_{\text{read}}(m_i)
\]

subject to

\[
\sum_{i=1}^{k} |m_i| \leq \sigma_{ss}
\]

\[
S_{\text{rep}}(m_i) \geq \sigma_{rep}
\]

\[
S_{\text{read}}(m_i) \geq \sigma_{read}
\]

\[
sim(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k]
\]
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \left\{ m_i \ldots m_k \right\} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{rep}(m_i) \geq \sigma_{rep} \]

\[ S_{read}(m_i) \geq \sigma_{read} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k] \]
Optimization Framework to capture compactness, representativeness & readability

Objective function: Optimize representativeness & readability scores

Ensure: summaries reflect key opinions & reasonably well formed

\[ M = \arg\max \left\{ m_i \ldots m_k \right\} \sum_{i=1}^{k} S_{\text{rep}}(m_i) + S_{\text{read}}(m_i) \]

\[ S_{\text{rep}}(m_i) \geq \sigma_{\text{rep}} \]

\[ S_{\text{read}}(m_i) \geq \sigma_{\text{read}} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{\text{sim}} \forall i, j \in [1, k] \]

2.3 very clean rooms
2.1 friendly service
1.8 dirty lobby and pool
1.3 nice and polite staff

Objective function value
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \{ m_i \ldots m_k \} \sum_{i=1}^{k} S_{\text{rep}}(m_i) \geq \sigma_{\text{rep}} \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{\text{read}}(m_i) \geq \sigma_{\text{read}} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{\text{sim}} \forall i, j \in [1, k] \]

**Constraint 1:** Maximum length of summary.
- User adjustable
- Captures compactness.
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \left\{ m_i \ldots m_k \right\} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma \]

\[ S_{rep}(m_i) \geq \sigma_{rep} \]

\[ S_{read}(m_i) \geq \sigma_{read} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k] \]

Constraint 2 & 3: Min representativeness & readability.
- Helps improve efficiency
- Does not affect performance
Optimization Framework to capture compactness, representativeness & readability

\[ M = \arg \max \left\{ m_i...m_k \right\} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma \]

\[ S_{rep}(m_i) \geq \sigma \]

\[ S_{read}(m_i) \geq \sigma \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{\text{sim}} \forall i, j \in [1, k] \]

**Constraint 4:** Max similarity between phrases
- User adjustable
- Captures **compactness** by minimizing redundancies
Optimization Framework to capture compactness, representativeness & readability

Ensures representativeness & readability

\[ M = \arg \max \{ m_i \ldots m_k \} \sum_{i=1}^{k} S_{rep}(m_i) + S_{read}(m_i) \]

subject to

\[ \sum_{i=1}^{k} |m_i| \leq \sigma_{ss} \]

\[ S_{rep}(m_i) \geq \sigma_{rep} \]

\[ S_{read}(m_i) \geq \sigma_{rep} \]

\[ \text{sim}(m_i, m_j) \leq \sigma_{sim} \forall i, j \in [1, k] \]
Now, we need to know how to compute:

- Similarity scores: \( \text{sim}(m_i, m_j) \)?
- Readability scores: \( \text{Sread}(m_i, m_j) \)?
- Representativeness scores: \( \text{Srep}(m_i, m_j) \)?
Similarity scoring, $\text{sim}(m_i, m_j)$

- Score similarity between 2 phrases
- User adjustable parameter

**Why important?**
- Allows user to control redundancy
- E.g. With *small devices* users may desire summaries with *good coverage* of information $\rightarrow$ *less redundancies*!

**Measure used:**
- Standard Jaccard Similarity Measure
- Can use other measures (e.g. cosine) – not our focus
Purpose: Measure how well a phrase represents opinions from the original text?

2 properties of a highly representative phrase:
1. Words should be strongly associated in text
2. Words should be sufficiently frequent in text

Captured by a modified pointwise mutual information (PMI) function
Representativeness scoring: $S_1(w_i)$

- Original PMI function, \( pmi(w_i, w_j) \):

\[
 pmi(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i) \times p(w_j)}
\]

Captures property 1: Measures strength of association between words.
Representativeness scoring, $S_{rep}(mi)$

- **Original PMI function, $pmi(w_i, w_j)$**
  
  $$pmi(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i) \times p(w_j)}$$

- **Modified PMI function, $pmi'(w_i, w_j)$**
  
  $$pmi'(w_i, w_j) = \log_2 \frac{p(w_i, w_j) \times c(w_i, w_j)}{p(w_i) \times p(w_j)}$$

  **Captures property 2:** Rewards well associated words with high co-occurrences

  **Add frequency of occurrence within a window**
To compute representativeness of a phrase:

\[
S_{rep}(w_1..w_n) = \frac{1}{n} \sum_{i=1}^{n} pmi_{lo}
\]

\[
pmi_{local}(w_i) = \left[ \frac{1}{2C} \sum_{j=i-C}^{i+C} pmi'(w_i, w_j) \right]
\]

Take average strength of association (pmi') of each word, wi in phrase with C neighboring words.
Representativeness scoring, $S_{rep}(mi)$

Gives a **good estimate** of how strongly associated the words are in a phrase.

To compute the representativeness of a phrase:

$$S_{rep}(w_1..w_n) = \frac{1}{n} \sum_{i=1}^{n} pmi_{local}(w_i)$$

$$pmi_{local}(w_i) = \left[ \frac{1}{2C} \sum_{j=i-C}^{i+C} pmi'(w_i, w_j) \right]$$
Readability scoring, $S_{read}(mi)$

- **Purpose:** Measure well-formedness of phrases
  - Phrases are constructed from seed words
    - can have new phrases not in original text
  - No guarantee phrases would be well-formed

- **Our readability scoring:**
  - Based on Microsoft's Web N-gram model
  - N-gram model used to obtain conditional probabilities of phrases
  
  $S_{read}(w_k...w_n) = \frac{1}{K} \log_2 \prod_{k=q}^{n} p(w_k | w_{k-q+1}...w_{k-1})$

  - **Intuition:** A phrase is more readable if it occurs more frequently on the web

  **chain rule** to compute joint probability in terms of conditional probabilities (averaged)
Example: Readability scores of phrases using tri-gram LM

<table>
<thead>
<tr>
<th>Ungrammatical</th>
<th>Grammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>“sucks life battery” -4.51</td>
<td>“battery life sucks” -2.93</td>
</tr>
<tr>
<td>“life battery is poor” -3.66</td>
<td>“battery life is poor” -2.37</td>
</tr>
</tbody>
</table>
We have all the scoring components...

- Scoring each candidate phrase is **not practical**!

- Why?
  - Phrases composed using words from original text
  - Potential **solution space** would be **Huge**!

- Our solution: **greedy summarization algorithm**
  - Explore solution space with **heuristic pruning**
  - Touch only the most **promising candidates**
Overview of summarization algorithm

Input

Text to be summarized

Unigrams

....
very
nice
place
clean
problem
dirty
room ...

Step 1: Shortlist high freq unigrams (count > median)

Seed Bigrams

very + nice
very + clean
very + dirty
clean + place
clean + room
dirty + place ...

Srep > σ_rep

Step 2: Form seed bigrams by pairing unigrams. Shortlist by S_{rep} \cdot (S_{rep} > \sigma_{rep})
Overview of summarization algorithm

Higher order n-grams

**Candidates** + **Seed Bi-grams** = **New Candidates**

- **very clean** + **clean rooms** = **very clean rooms**
- **very dirty** + **dirty room** = **very dirty room**
- **very nice** + **nice place** = **very nice place**

**Step 3: Generate higher order n-grams.**
- Concatenate existing candidates + seed bigrams
- Prune non-promising candidates ($S_{rep}$ & $S_{read}$)
- Eliminate redundancies ($sim(mi,mj)$)
- Repeat process on shortlisted candidates (until no possibility of expansion)

**Summary**

- 0.9  very clean rooms
- 0.8  friendly service
- 0.7  dirty lobby and pool
- 0.5  nice and polite staff

**Sorted Candidates**

$S_{rep}<\sigma_{rep}; S_{read}<\sigma_{read}$

**Step 4: Final summary.**
Sort by objective function value. Add phrases until $|M|<\sigma_{ss}$
Evaluation...
Product reviews from **CNET** (330 products)

- Each product has minimum of 5 reviews

*Content Summarized*

"Best Phone I've ever owned.....AMAZING Battery Life!!!!" on February 4, 2012 by XXXXXX

**Pros:**
- Battery Life (See Below)
- Display
- Battery Life
- Functionality
- Battery Life
- Smart Actions Battery Life Saver

**Cons:**

- It wasn't as intuitive or reactive as the iPhone, but I want my phone to be a phone. If I want the crisperst action shots, I'll use my camera.

**Summary:**
I have the Droid X2 and then upgraded to the 16gb Razr. The Razr is the exact same phone as the Maxx, but my battery was dying every day before I left the office. Don't undervalue how much battery life 4gLTE service actually uses. I used Smart Actions, made my own adjustments, limited use, didn't read my Kindle app at lunch, and still had a dead battery on my way home from work. So, I traded up to the Razr Maxx. This week, I've used my phone to watch
Summarization Task

Given

Textual reviews about a product

Task

Micropinion Summary, \( M \)

\( m_1 \)
2.3 easy to use

\( m_2 \)
2.1 lense is not clear

\( m_3 \)
1.8 too big for pocket

\( m_k \)
1.3 expensive batteries

\[ \text{argmax} \ S_{\text{rep}}(m_i) + S_{\text{read}}(m_i) \]

Constraints

summary size, \( \sigma_{\text{ss}} \)

redundancy, \( \sigma_{\text{sim}} \)

representativeness, \( \sigma_{\text{rep}} \)

readability, \( \sigma_{\text{read}} \)
Gold Standard

- **Human composed summaries**
  - Two human summarizers
    - Each summarize 165 product reviews (total 330)
  - Top 10 phrases from **pros & cons** provided as **hints**
    - Hints help with topic coverage ➔ reduce bias
  - Summarizers asked to compose a **set of short phrases** (2-5 words) summarizing **key opinions** on the product
3 representative baselines:

- **Tf-Idf** – *unsupervised* method commonly used for key phrase extraction tasks
  - Selected only *adjective* containing n-grams (performance reasons)
  - *Redundancy removal* used to generate non-redundant phrases

- **KEA** – state of the art *supervised* key phrase extraction model [witten et al. 1999]
  - Uses a *Naive Bayes* model
  - Trained using 100 review documents withheld from dataset

- **Opinosis** – *unsupervised* abstractive summarizer designed to generate textual opinion summaries [ganesan et al. 2010]
  - Shown to be effective in generating *concise opinion summaries*
  - Designed for highly *redundant* text
Quantitative Evaluation

- **ROUGE** - to determine quality of summary (ROUGE-1 & ROUGE-2)
  - **Standard measure** for summarization tasks
  - Measures **precision** & **recall** of overlapping units between **computer generated** summary & **human** summaries
Qualitative Evaluation

- **Questionnaire** – to assess potential utility of generated summaries to users
  - Answered by 2 assessors
  - Original reviews provided as reference

**Questionnaire**


- **Grammaticality** [DUC 2005]
  - Are the phrases readable?

- **Non-redundancy** [DUC 2005]
  - Are the phrases in the summary unique?

- **Informativeness** [Filippova 2010]
  - Do the phrases convey **important** information about the product?
Results...

- Our approach is referred to as WebNGram
Performance comparisons

- **WebNgram**: Performs the best for this task
- **Opinosis**: much better than KEA & tfidf
- **KEA**: slightly better than tfidf
- **Tfidf**: Worst performance

Summary Size (max words) vs ROUGE-2 RECALL
# of Generated Phrases

- **Intuition:** Well-formed phrases tend to be longer in general

  - *very clear screen* vs. *very clear*
  - *good battery life* vs. *good battery*
  - *screen is bright* vs. *screen is*

- A few longer phrases is more desirable than many fragmented (i.e. short) phrases
# of Generated Phrases

**KEA**: Generates most # of phrases (i.e. favors short phrases)

**WebNGram**: Generates fewest phrases on average. (each phrase is longer)

WebNGram phrases are generally more well-formed
In our algorithm: n-grams are generated from seed words
- potential of forming new phrases not in original text

Why not use existing n-grams?
- With redundant opinions, using seen n-grams may be sufficient
- Performed a run by forcing only seen n-grams to appear as candidate phrases.
Our search algorithm helps discover useful new phrases.
Example:

Unseen N-Gram (Acer AL2216 Monitor)

“wide screen \textit{lcd monitor is bright}”
readability : -1.88
representativeness: 4.25

“...plus the \textit{monitor} is very \textit{bright}...”
“...it is a \textit{wide screen}, great color, great quality...”
“...this \textit{lcd monitor} is quite \textit{bright} and clear...”
Stability of Non-User Dependent Parameters, $\sigma_{\text{rep}}$ & $\sigma_{\text{read}}$

- **Purpose of $\sigma_{\text{rep}}$ & $\sigma_{\text{read}}$:**
  - Control minimum representativeness & readability
  - Helps *prune* non-promising candidates $\implies$ improves efficiency of algorithm

- **Without $\sigma_{\text{rep}}$ & $\sigma_{\text{read}}$ -** we would still arrive at a solution, however:
  - Time to convergence would be much *longer*
  - Results could be *skewed*
  - These parameters need to be set correctly!
Stability of Non-User Dependent Parameters, $\sigma_{\text{rep}}$ & $\sigma_{\text{read}}$

Performance is **stable** except in extreme conditions - thresholds are **too restrictive**
Stability of Non-User Dependent Parameters, $\sigma_{\text{rep}}$ & $\sigma_{\text{read}}$

ROUGE-2 scores at various $\sigma_{\text{read}}$ settings

ROUGE-2 scores at various $\sigma_{\text{rep}}$ settings

Ideal setting:
- $\sigma_{\text{read}}$: between -2 and -4
- $\sigma_{\text{rep}}$: between 1 and 4
Manual assessment of summaries

**Questionnaire** [Score 1-5]
- Score < 3 → poor
- Score > 3 → good

**Grammaticality** [DUC 2005]
- WebNgram: 4.2/5
- TfIDF: 2.0/5
- Human: 4.7/5

**Non-redundancy** [DUC 2005]
- WebNgram: 3.9/5
- TfIDF: 2.3/5
- Human: 4.5/5

**Informativeness** [Filippova 2010]
- WebNgram: 3.2/5
- Tfidf: 1.7/5
- Human: 3.6/5
Manual assessment of summaries

Questionnaire [Score 1-5]
Score < 3 ➔ poor
Score > 3 ➔ good

Grammaticality [DUC 2005]
- WebNgram: 4.2/5
- TfIDF: 2.0/5
- Human: 4.7/5

Non-redundancy [DUC 2005]
- WebNgram: 3.9/5
- TfIDF: 2.3/5
- Human: 4.5/5

Informativenesseness [Filippova 2010]
- WebNgram: 3.2/5
- TfIDF: 1.7/5
- Human: 3.6/5

TFIDF: Poor scores on all 3 aspects (all below 3.0)
Manual assessment of summaries

**Questionnaire** [Score 1-5]
- Score < 3 ➔ poor
- Score > 3 ➔ good

**Grammaticality** [DUC 2005]
- WebNgram: 4.2/5
- TfIDF: 2.0/5
- Human: 4.7/5

**Non-redundancy** [DUC 2005]
- WebNgram: 3.9/5
- TfIDF: 2.3/5
- Human: 4.5/5

**Informativeness** [Filippova 2010]
- WebNgram: 3.2/5
- TfIDF: 1.7/5
- Human: 3.6/5

**WebNNGram:** All scores above 3.0
Close to human scores
Manual assessment of summaries

Questionnaire [Score 1-5]
Score < 3 ➔ poor
Score > 3 ➔ good

- Grammaticality [DUC 2005]
  - WebNgram 4.2/5
  - Tfidf 2.0/5
  - Human 4.7/5

- Non-redundancy [DUC 2005]
  - WebNgram 3.9/5
  - Tfidf 2.3/5
  - Human 4.5/5

- Informativeness [Filippova 2010]
  - WebNgram 3.2/5
  - Tfidf 1.7/5
  - Human 3.6/5

Subjective aspect. Human summaries scored slightly better than WebNgram
Known Issues

- Semantic redundancies
  "Good sound quality" ≠ "Excellent audio" ➞ semantically similar
  - Due to directly using surface words
  - Solution: opinion/phrase normalization before summarizing

- Some phrases are not opinion phrases
  "I bought this for Christmas" ➞ grammatical & representative
  - Due to our candidate selection strategy
    - Plus Side: Very general approach, can summarize any text
    - Down Side: Summaries may be too general
  - Solution: stricter selection of phrases
    - E.g. Select only opinion containing phrases
A Sample Summary

Canon Powershot SX120 IS

Easy to use
Good picture quality
Crisp and clear
Good video quality

Useful for pushing opinions to devices where the screen is small
Summary

- Optimization framework to generate **ultra-concise summaries** of opinions
  - **Emphasis:** representativeness, readability & compactness

- Evaluation shows our summaries are:
  - Well-formed and convey essential information
  - More effective than other competing methods

- Our approach is **unsupervised, lightweight & general**
  - Can summarize any other textual content (e.g. news articles, tweets, user comments, etc.)
Thanks! Questions?