CSCE 771: Computer Processing of Natural Language
Lecture 4: NLP Tasks, NLP for Business

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

30TH SEPTEMBER, 2022

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 4

- Opening Segment
  - Recap of Lecture 3
  - Class rules

- Main Lecture

- Concluding Segment
  - Course project
  - About next lecture – Lecture 5

Main Section
- NLP Tasks
- Case Study: NLP for Business
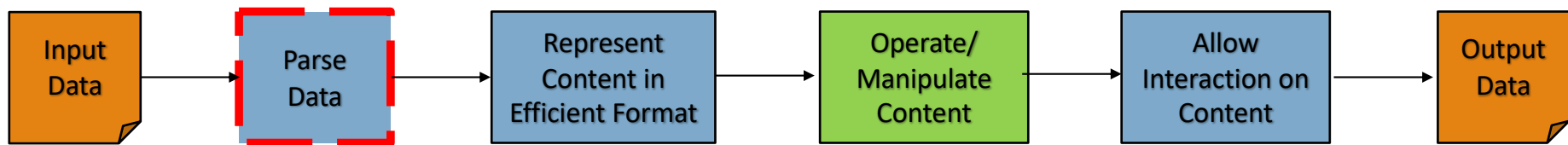
Recap of Lecture 3

- We
  - Looked at structure of building block of text, i.e., words, from an English and non-English perspective
  - Reviewed morphology
  - Looked at useful lexicons that can help simplify language tasks

- Read paper #1-1 - *Contextual Word Representations: Putting Words into Computers* (Full list: https://github.com/biplav-s/course-nl-f22/tree/main/reading-list)

- Contranyms
  - Sanction : allow / prevent
  - Bomb : success / failure
  - Break : opportunity / problem
  - Dust: remove or put dust
  - More examples: https://en.wiktionary.org/wiki/Appendix:English_contranyms

Class Rules

- Absence policy
  - Excused absence: If you have a reason to miss, please send me a note BEFORE the class.
  - Non-excused absence: Will be noted in the excel sheet: https://docs.google.com/spreadsheets/d/1xBOs4loR_iagqzj6Eo4vrhpY3Mm1e_smXXN3Q8Q5vIg/edit?usp=sharing in column E.
  - More than two non-excused absences will lead to a 10% penalty on cumulative score.

- Honor code and plagiarism policy
  - Go through: https://sc.edu/about/offices_and_divisions/student_conduct_and_academic_integrity/instructors/promoting_academic_integrity/index.php
  - On **plagiarism**
    - Acknowledging / giving credits is essential when reusing material
    - Your work cannot be complete reuse by someone else; especially project
    - Will lead to being reported to corresponding office and a D or failed grade

Main Lecture

NLP Basic Tasks

Input Data → Parse Data → Represent Content in Efficient Format → Operate/ Manipulate Content → Allow Interaction on Content → Output Data

Tasks Depends on Applications

- Tokenization – getting tokens for processing

- Normalization - making into canonical form

- Case folding – handling cases

- Lemmatization – handling variants (shallow)

- Stemming – handling variants (deep)

Task - Tokenization: Some Examples

- Example 1
  - **Input**: 'At eight o\'clock on Thursday morning, we will go to school. There is' + \
          'No one there but myself.
  - '**Output**: ['At', 'eight', "o'clock", 'on', 'Thursday', 'morning', ',', 'we', 'will', 'go', 'to', 'school', '.', 'There', 'isNo', 'one', 'there', 'but', 'myself', '.']

- Example 2
  - **Input**: "CSCE 771: Computer Processing of Natural Language " + \
    " Lecture 3: Words, Morphology, Lexicons" + \
    " Prof. Biplav Srivastava, AI Institute   31st Aug 2020 "'
  - **Output**: ['CSCE', '771', ':', 'Computer', 'Processing', 'of', 'Natural', 'Language', 'Lecture', '3', ':', 'Words', ',', 'Morphology', ',', 'Lexicons', 'Prof.', 'Biplav', 'Srivastava', ',', 'AI', 'Institute', '31st', 'Aug', '2020']

- Comments:
  - Does not capture compound words

Tokenization Issue and Practices

- **Penn Treebank** Tokenization
  - Separates clitics - doesn't becomes does plus n't
  - Keeps hyphenated words together
  - Separates out all punctuation

```
Input:   "The San Francisco-based restaurant," they said,
         "doesn't charge $10".
Output:  " The San Francisco-based restaurant , " they said ,
         " does n't charge $ 10 " .
```

- Ambiguity while tokenization
  - the book's cover => the book's cover
  - they're => they are

- Practice
  - Deterministic algorithms based on regular expressions compiled into very efficient finite state automata
  - Breaks ambiguity along conventions

Source: Jurafsky & Martin

Tokenization in Other Languages

- French
  - *L'ensemble* → one token or two?
    - *L* ? *L'* ? *Le* ?
    - Want *l'ensemble* to match with *un ensemble*

- German noun compounds are not segmented
  - *Lebensversicherungsgesellschaftsangestellter*
  - 'life insurance company employee'
  - German information retrieval needs **compound splitter**

Source: Jurafsky & Martin

Tokenization in Other Languages

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida

- Japanese has alphabets intermingled from other "languages"
  - Dates/amounts in multiple formats

フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

Katakana   Hiragana   Kanji   Romaji

Word Tokenization in Chinese

Also called **Word Segmentation**

Chinese words are composed of characters
◦ Characters are generally 1 syllable and 1 morpheme.
◦ Average word is 2.4 characters long.

Standard baseline segmentation algorithm:
◦ Maximum Matching (also called Greedy)

Source: Jurafsky & Martin

(2.5) 姚明　进入　总决赛
　　　　YaoMing reaches finals

or as 5 words ('Peking University' segmentation):

(2.6) 姚　明　进入　总　决赛
　　　　Yao Ming reaches overall finals

Finally, it is possible in Chinese simply to ignore words altogether and use characters as the basic elements, treating the sentence as a series of 7 characters:

(2.7) 姚　明　进　入　总　决　赛
　　　　Yao Ming enter enter overall decision game

Task - Normalization

- Have words/ word phrases in canonical form

- Examples
  - **UoSC** and **USC**
  - *U.S.A.* and *USA*
  - **Window** and **windows**
  - **"Dr. Martin Luther King Jr."** and **"Dr. M. L. King"**

- Why do it
  - Improve match between query and index
  - Recognize duplicates

- Standards
  - Physical Address - USPS

Task - Case Folding

- Handle case(s) of letter(s)

- Examples
  - Force words to lower case
  - Retain cases of words as entered
    - **US** versus **us** is important

- Case can have important meaning
  - Acronym
  - Useful in sentiment analysis

Task - Parts of Speech (POS) Tagging with Homonym

Data

"They refuse to permit us to obtain the refuse permit"

POS

[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), (**'permit'**, 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), (**'permit'**, 'NN')]

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Table Source:
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Task - Parts of Speech (POS) Tagging

Data
"Hola Class! Dost kya kar rahe ho? Let us pay attention."

POS
[('Hola', 'NNP'), ('Class', 'NN'), ('!', '.'), ('Dost', 'NNP'), ('kya', 'VBD'),
('kar', 'JJ'), ('rahe', 'NN'), ('ho', 'NN'), ('?', '.'), ('Let', 'VB'), ('us', 'PRP'),
('pay', 'VB'), ('attention', 'NN'), ('.', '.')]

Table Source:
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Task - Lemmatization

- Reduce inflections or variant forms to base form
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- Examples
  - churches : church
  - rocks : rock
  - children : child
  - focii : focii

See sample code on how to perform

Code Sample - NLTK

Notebook:

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l4-nlp-basictasks/NLTK%20Basic%20Tasks.ipynb

Task - Stemming

- Reduce terms to their stems in information retrieval

- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., **automate(s), automatic, automation** all reduced to **automat**.

Complexity
- Multiple senses of words
  - Sense ambiguities: **ceil-** is not the stem of **ceiling**
- Complicated morphological rules
- Part-of-speech
- Irregular words

for example compressed and compression are both accepted as equivalent to compress.

➡

for exampl compress and compress ar both accept as equival to compress

See sample code on how to perform

Complexity of Stemming

- Why affixes are added: Grammatical role, tense, derivational morphology

- Complexity reasons
  - Multiple senses of words
    - Sense ambiguities: *ceil-* is not the stem of *ceiling*
  - Complicated morphological rules
  - Part-of-speech
  - Irregular words

Source: NLTK Documentation

Lemmatization and/ or Stemming

- Lemmatization is usually a good idea; reduces variants

- Stemming can cause issues – over and under stemming
  - Use it when doing careful word analysis

- Issue more complicated when doing multi-lingual analysis
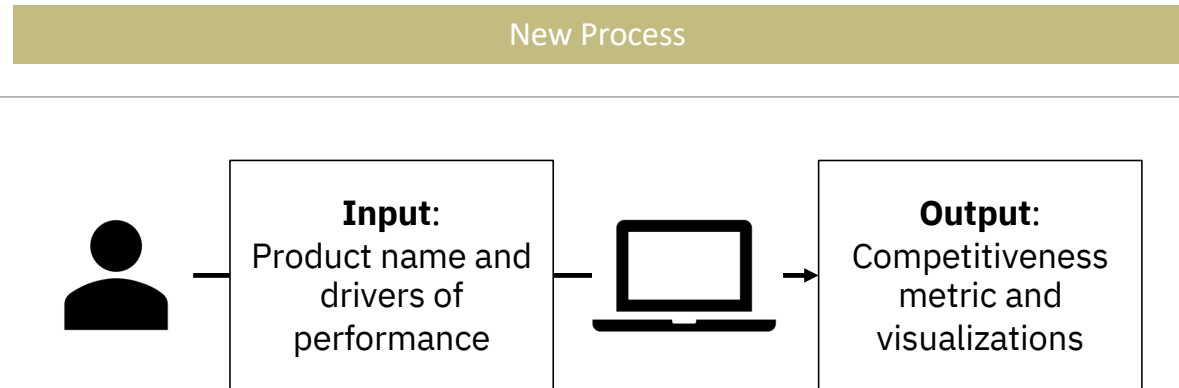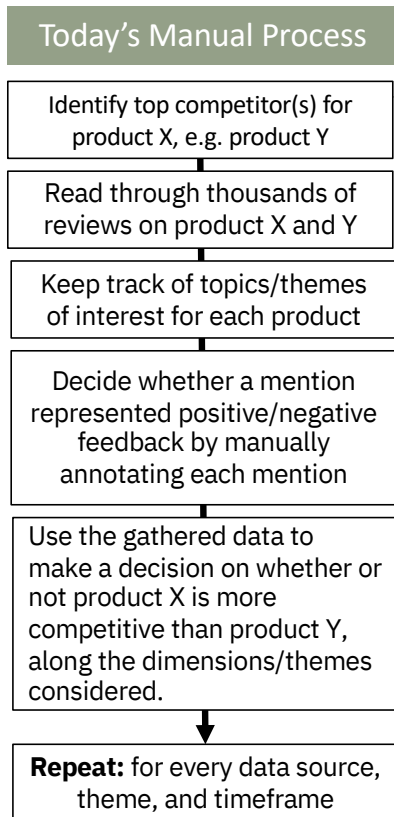
Code Sample - Spacy

Notebook:

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l4-nlp-basictasks/Spacy%20Basic%20Tasks.ipynb

Case Study: NLP for Business - Market Intelligence

Clarity: Data-Driven Competitive Analysis

1. Sheema Usmani, Mariana Bernagozzi, Yufeng Huang, Michelle Morales, Amir Sabet Sarvestani, Biplav Srivastava, Clarity: Data-driven Automatic Assessment of Product Competitiveness, IAAI/AAAI 2020, **Deployed Application Award**
2. (Demo paper) Data-driven ranking and visualization of products by competitiveness, Sheema Usmani, Mariana Bernagozzi, Yufeng Huang, Michelle Morales, Amir Sabet Sarvestani, Biplav Srivastava, AAAI 2020
3. Yufeng Huang, Mariana Bernagozzi, Michelle Morales, Sheema Usmani, Biplav Srivastava, Michelle Mullins, Clarity 2.0: Improved Assessment of Product Competitiveness from Online Content. AI Mag. 42(2): 59-70 (2021)

Competitive Analysis: Before & After

Today's Manual Process

New Process

Today's Manual Process

- Identify top competitor(s) for product X, e.g. product Y
- Read through thousands of reviews on product X and Y
- Keep track of topics/themes of interest for each product
- Decide whether a mention represented positive/negative feedback by manually annotating each mention
- Use the gathered data to make a decision on whether or not product X is more competitive than product Y, along the dimensions/themes considered.
- **Repeat:** for every data source, theme, and timeframe

Input: Product name and drivers of performance

Output: Competitiveness metric and visualizations

Steps

1. Prepare review data of products p_1 to p_N from sources d_1 to d_M (offline)
2. Process request for analysis for product p_i (online)
3. Visualize analysis results (online, optional)

Illustrative Output

Video demo
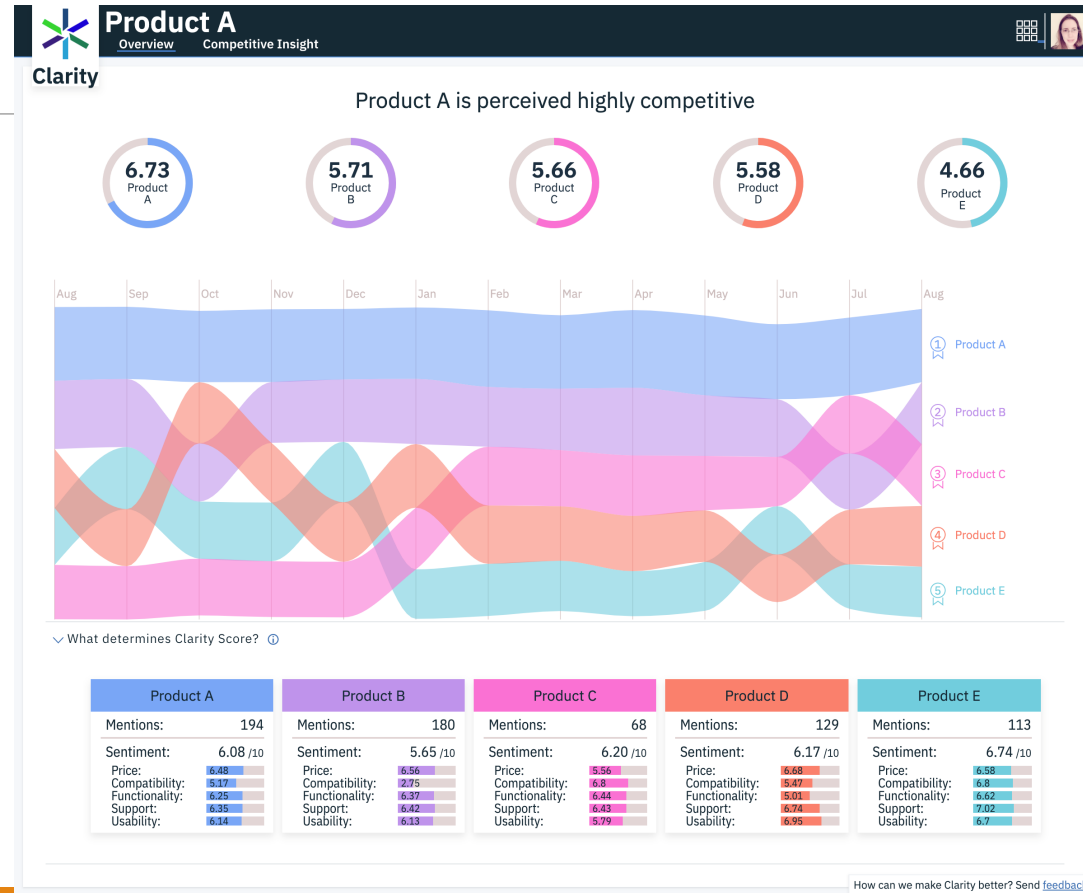
Clarity Score and Trends

Drivers and Raw Scores

Evaluation and Impact

- The system has been running for over two years and used by over 4500 people performing over 200 competitive analyses involving over 1000 products Clarity has been running

- In-field evaluation
  - High user satisfaction
    - Net Promoter Score (NPS) of 52; Scale -100 to 100

- In-lab evaluation
  - Scores consistent with Gartner's Magic Quadrant
    - Products v/s Vendor ranking
  - Clarity scores consistent with Net Promoter Score (NPS) of 50 products
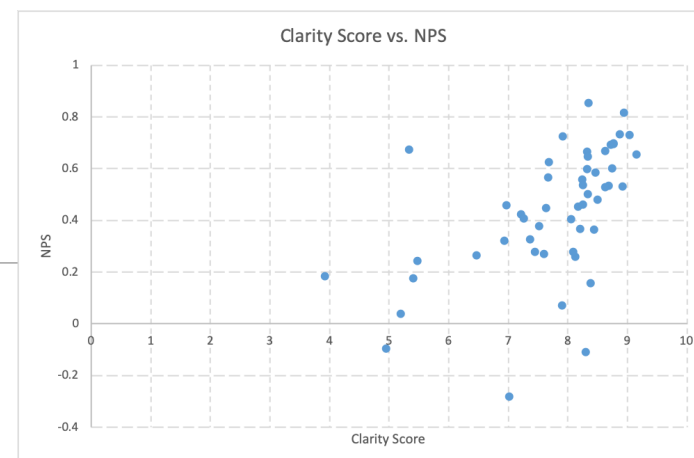
$$\mathcal{CS}(p_L) > \mathcal{CS}(p_C) > \mathcal{CS}(p_N)$$

$$\mathcal{CS}(p_L) > \mathcal{CS}(p_V) > \mathcal{CS}(p_N)$$

Clarity Score vs. NPS
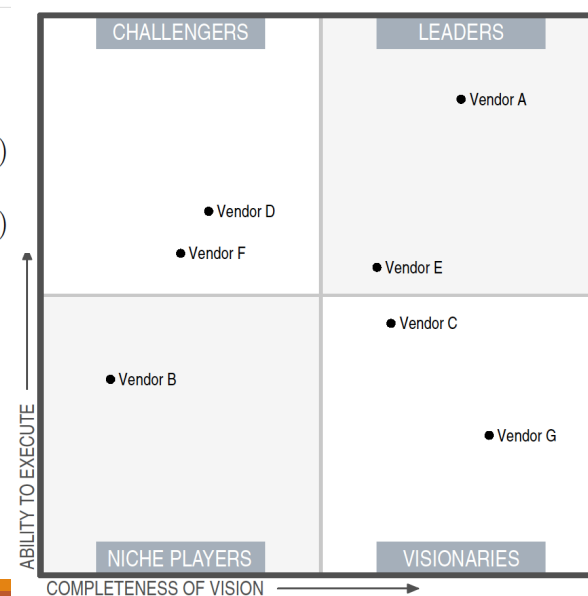
Clarity 2.0

We still consider:
- the sentiments on different topics
- the overall rating,
- and the content volume

But change calculation from statistical measures to Bayesian statistics

Model	Precision	Recall	F1 Score
Neural network on sentence	0.40	0.32	0.36
Neural network on words	0.56	0.40	0.47
Neural network based on attention	0.56	0.56	0.56
Human performance	0.75	0.56	0.64

Table 1: Model performance on one of the topics, *product features and functionalities*. Precision, recall, and F1 score are compared for the models described in the *Improved NLP Model in Clarity 2.0* section. It should be noted that the human performance is relative low compared to typical academic datasets. The reason is that practical NLP applications are typically ambiguous, which results in more noise in the results and leads to the overall lower score compared to well-refined academic datasets.

Much easier to use than the competition and you will see nearly immediate results . (Ease of use - 0.98308)
Much easier to use than the competition and you will see nearly immediate results . (Performance and efficiency - 0.53858)
Much easier to use than the competition and you will see nearly immediate results . (Product competitiveness - 0.50202)

Figure 6: Example of topic classification using the attention mechanism. The example sentence is classified into three topics, *ease of use*, *performance and efficiency*, and *product competitiveness*, with corresponding probabilities, respectively. The darkness of the highlighted keyword is calculated by multiplying the attention weight with the classification probability, which reflects the contribution of the keyword to the classification of a certain topic.
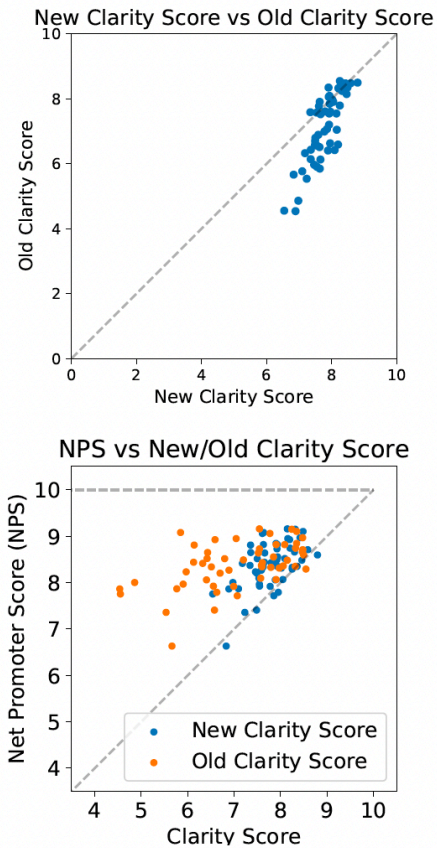
Improved Performance

Figure 7: Comparison between NPS, *Clarity Score 1.0*, and *Clarity Score 2.0*. Top: comparison between the old and new *Clarity Scores*. Bottom: comparison between the *Clarity Scores* and NPS.

Yufeng Huang,
Mariana Bernagozzi,
Michelle Morales,
Sheema Usmani,
Biplav Srivastava,
Michelle Mullins,
Clarity 2.0: Improved
Assessment of
Product
Competitiveness
from Online Content.
AI Mag. 42(2): 59-70
(2021)

Market	Vendor	Product	Gartner Q	*Clarity* score
M_1 = Data Science Machine Learning Platform	$V_1^{M_1}$	$P_1^{V_1 M_1}$	Visionary	8.79
	$V_2^{M_1}$	$P_1^{V_2 M_1}$	Visionary	8.48
	$V_3^{M_1}$	$P_1^{V_3 M_1}$	Challenger	7.93
	$V_4^{M_1}$	$P_1^{V_4 M_1}$	Challenger	7.91
	$V_5^{M_1}$	$P_1^{V_5 M_1}$	Leader	7.34
M_2 = Data Management Solutions for Analytics	$V_1^{M_2}$	$P_1^{V_1 M_2}$	Leader	8.58
	$V_2^{M_2}$	$P_1^{V_2 M_2}$	Leader	8.46
	$V_3^{M_2}$	$P_1^{V_3 M_2}$	Leader	8.19
	$V_4^{M_2}$	$P_1^{V_4 M_2}$	Leader	7.96
	$V_5^{M_2}$	$P_1^{V_5 M_2}$	Leader	7.88
M_3 = Data Management Solutions for Analytics	$V_1^{M_3}$	$P_1^{V_1 M_3}$	Leader	7.62
	$V_2^{M_3}$	$P_1^{V_2 M_3}$	Leader	7.53
	$V_3^{M_3}$	$P_1^{V_3 M_3}$	Leader	7.36
M_4 = Operational Database Management Systems	$V_1^{M_4}$	$P_1^{V_1 M_4}$	Leader	8.58
	$V_2^{M_4}$	$P_1^{V_2 M_4}$	Leader	8.46
	$V_3^{M_4}$	$P_1^{V_3 M_4}$	Leader	8.19
M_5 = Analytics and Business Intelligence Platforms	$V_1^{M_5}$	$P_1^{V_1 M_5}$	Leader	8.79
	$V_1^{M_5}$	$P_2^{V_1 M_5}$	Leader	8.25
	$V_2^{M_5}$	$P_1^{V_2 M_5}$	Niche Player	7.92
	$V_3^{M_5}$	$P_1^{V_3 M_5}$	Niche Player	7.85

Table 2: Comparison of ratings provided by Gartner vs. *Clarity Score*

Lecture 4: Concluding Comments

- We looked at a variety of NLP tasks

- We looked at evaluation methods
  - For NLP tasks (Precision, Accuracy, Recall)
  - For ML tasks (AUC)

Concluding Segment

Choosing a Project – Some Considerations

- Scope: what is the problem?

- Current-state: what happens in the problem today?

- Who cares: who will benefit with the problem being solved?

- Desired-state: what will be the future situation if your project succeeds?

- Resources/ dataset: do you have reasonable data and compute resources to do the work?

- Evaluation: how will we measure goodness of the work?

Review project spreadsheet

Discussion: Course Project

- Expectations
  - Apply methods learned in class or of interest to a problem of interest
  - Be goal oriented: aim to finish, be proactive, be innovative
  - Do top-class work: code, writeup, presentation

- Typical pitfalls
  - Not detailing out the project, assuming data
  - Not spending enough time

Review Current List of Selections

- What will be awarded
  - Results and efforts (balance)
  - Challenge level of problem

Discussion: Reading Material

- **Paper**: "Contextual Word Representations: Putting Words into Computers", by Noah Smith, CACM June 2020

- Key Points
  - Different representations of words
    - Characters, one-hot encoding, vectors
  - Determining contexts of words is important
    - Contextual word vectors
    - Bi-, tri-, N-gram representations

About Next Lecture – Lecture 5

Lecture 5:

- Discussion on reading material - **Paper:**
 "Contextual Word Representations: Putting Words into Computers", by Noah Smith,
  CACM June 2020

- Parsing

- Projects
  - Review project topics in class