A NLP BUSINESS PROJECT

Hanhan Wu

ABOUT ME

- SFU Big Data
- Vancity data science intern
- Side Projects
  - Mining Social Media
  - NLP Practice
  - Data Preprocessing & Analysis
  - Big Data Development
  - ML Algorithms Implementation

Typical Foodie

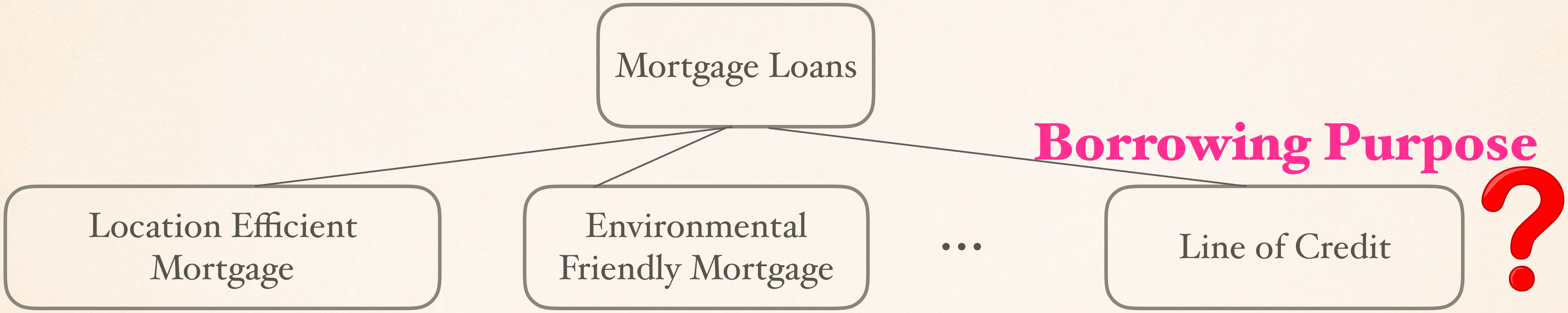
PURPOSE

Asking for **your feedback**, to help further NLP work

LITTLE ABOUT VANCITY

- Vancity Savings and Credit Union
- Savings and Lending Business
- No customer, but "member"
- People-Planet-Profit $\longrightarrow$ Impact

PROJECT PURPOSE



Lender Notes Mock-up Sample

- "both members are new to Vancity.Member is purchasing a revenue property at Redmond St NO. 999. for $800,000. Total financing $123,456. Member's husband will go on at a covanator on the mortgageas member is on maternity leave."
- "mbr is looking at purchasing a home located at Redmond St.vancouver bcMay.12.2013material change to show credit card as POM as well as to get appraisal review as LOS skipped this step. no other changes have been made.Previous approval has expired, member has found a new property she has an accepted offer on, amount lower than previous approval.amount reduced from 752k to 750k, also changed from purchase to refinance as member has decided to use some of her stocks to purchase the property and then will use this credit line to repurchase the stock as per her accountants suggestion."

METHODS TRIED

- **Topic Extraction** with Spark ML, Scikit-Learn
- **Find Optimal Cluster Numbers** with R text mining package
- **Pattern Extraction** with Self-implemented methods
- Smart **UW NLP** algorithms
- **Key Components Search** with Self-implemented methods

TOPIC EXTRACTION

Purpose Raw Text → Find →

Borrowing Purpose:
- Purchasing Mortgage?
- Pay other debt?
- ......?

Topic Extraction?

Basic Data Cleaning

- all to lower case
- remove special characters

→

- NMF
- LDA

Not persuasive to the business audience

NMF Sample Output

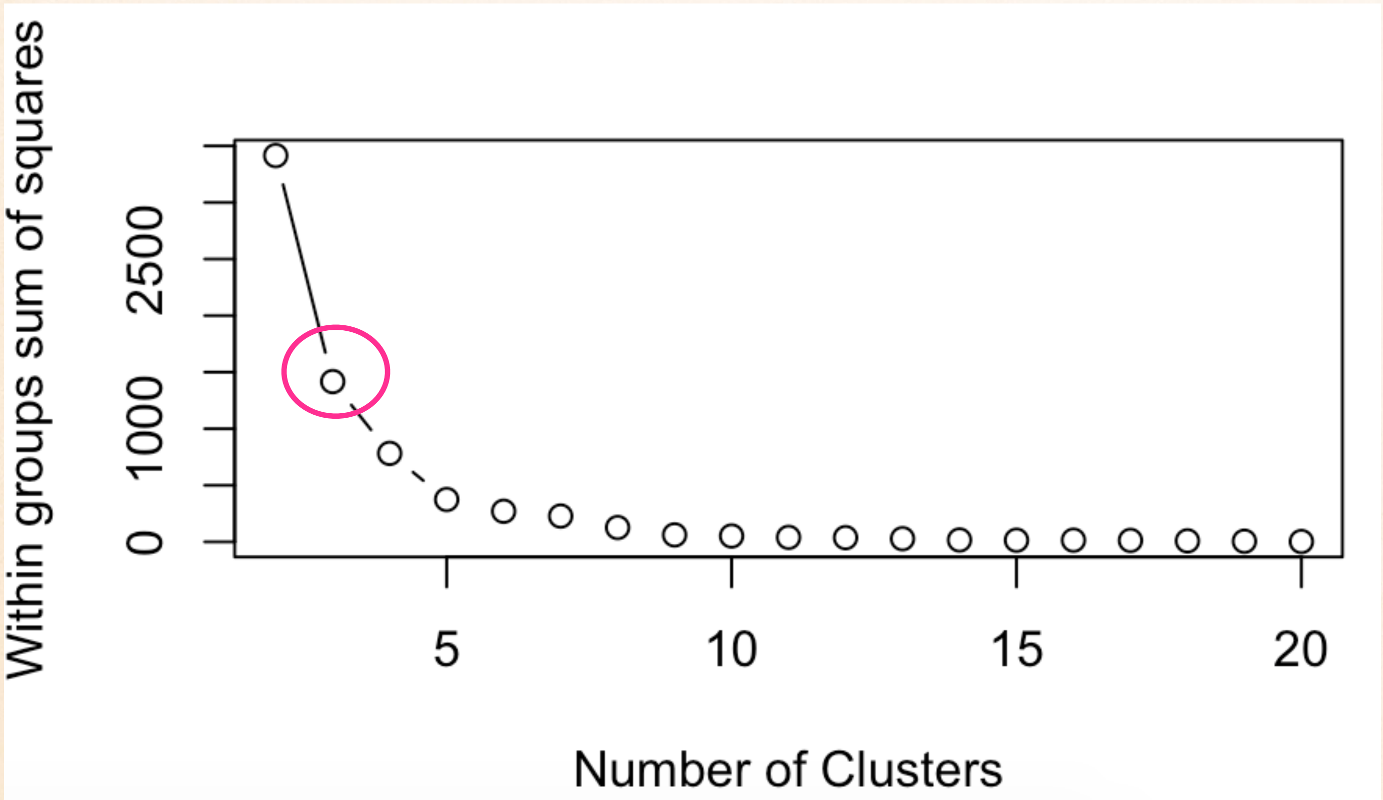
```
Topics in NMF model:
Topic #0:
member property mortgage purchase new purpose mtg creditline credit loc existing like revenue vancity members
```

LDA Sample Output

```
Topics in LDA model:
Topic #0:
safescan address fraud opened sin file identified employer cb member dob indications appraisal 1992
```

OPTIMAL CLUSTER NUMBERS

Maybe I should find the **optimal number of topics** first?



- **3 clusters** is the optimal , since SSE dropped dramatically

- But the topic extraction output still cannot be persuasive to the business audience

PATTERN EXTRACTION - PATTERNS

- Here, **NN** can be a continuous list of tokens start with 'NN' tag, such as NN, NNS, NNP or NNPS
- **VB** can be tokens start with 'VB' tag, such as VB, VBD, VBG, VBN, VBP or VBZ

Penn Treebank P.O.S tags	Description/Example
MD + TO + VB + NN	• **Example**: "would like to" + VB + NN
"for" + NN	• I changed the tag of "for" as 'for'
"member"/"members"/"mbr" + VB ... NN ... '.'/':'	• The sentences start with "member" or "members" or "mbr" • **...** here means unnecessary parts • The sentences end with tokens tagged as '.' or ':'
TO + VB + ... '.'/':'	• **...** here will all be included • The sentences end with tokens tagged as '.' or ':'
"want"/"wants" + TO + VB ... NN ... '.'/':'	• The sentences start with "want" or "wants" • **...** here means unnecessary parts • The sentences end with tokens tagged as '.' or ':'
VB + NN	• VB + continuous list of NN (for comparison)
NN	• Continuous list of NN (for comparison)

PATTERN EXTRACTION - EXTRACTION

Data Preprocessing

- Pick out special words such as “Vancity”, other words all convert **to lowercase**
- **Remove** html tags and **special characters**
- **Separate** “sentence_1.sentence_2” in to 2 **sentences**
- **Tokenize sentences** so that each word becomes (word, tag) format
- No stemming is better here, otherwise, token tags are less accurate

Output Sample

```
(u'for emergency', 15)
(u'for investment', 15)
(u'for renovations', 9)
(u'for mortgage', 8)
(u'for possible future investment', 6)
(u'for appliction', 6)
(u'for purpose', 6)
(u'for home', 5)
(u'for expenses', 4)
(u'for payment', 4)
(u'for estate', 4)
(u'for financing', 4)
(u'for refinance', 3)
(u'for future expenses', 3)
(u'for renewal', 3)
```

Check in further detail, what kind of investments

Output Sample

```
(u'credit bureau', 22)
(u'revenue property', 13)
(u'revenue properties', 10)
(u'takeout applications', 8)
(u'equity takeout applications', 8)
(u'crl mtg', 7)
(u'investment opportunities', 7)
(u'bypass operation', 4)
(u'term mortgage', 4)
(u'home renovations', 4)
(u'safety netaml', 3)
(u'term mtg', 3)
(u'kids education', 3)
```

Generate impact

UNIVERSITY of WASHINGTON

REVERB & OLLIE

- **ReVerb**
  - No need to clean data
  - Find binary relationship from the text
  - **Example Output**

```
"Funds will be used to invest in other channels"
"member wants to have funds"
```

- **Ollie**
  - No need to clean data
  - Find binary relationship from the text
  - Will find the relationship missed by ReVerb, such as longer relations
  - **Example Output**

```
"(member; plan to; renovate home;)"
"(she; want to invest in; education;)"
```

KEY COMPONENTS SEARCH

Use Case

Raw Text

- ... for investment opportunities
- ... for future investment
- ... invest on future mortgage opportunity
- ... in case of emergency or investment opprtunity miss-spelling
- ......

Want to know further about **investment opportunities** → Search → "Investment Opportunity"

Methods for Implementation

- Stemming is important here
- Calculation 1 - Calculate query words distance score
  - **Example**: "invest opportunity" vs. "invest.... opportunity", choose the closest one
- Calculation 2 - Calculate query words location score
  - **Example**: appears closer to the head of the paragraph, get higher score
- Calculation 3 - Calculate query words frequency score
  - **Example**: higher frequency, higher score
- Normalize scores from calculation 1, 2 & 3, give weights and combine the scores together

  In this case, **query words distance plays a more significant role**

THANK YOU & FEEDBACK

- **Topic Extraction** with Spark ML, Scikit-Learn
- **Find Optimal Cluster Numbers** with R text mining package
- **Pattern Extraction** with Self-implemented methods
- Smart UW NLP algorithms
- **Key Components Search** with Self-implemented methods