Email Summarize

Zihang Li, Yangpan Tao, Chao Wang

Design Procedure

- Use jape rules and gazetteers to get rid of noise information in documents

- Get the TF-IDF value for every worlds and save the top 10 words to a gazetteer

- Run Gate one more time to find the Noun Phase that contains the top 10 TF-IDF words

Annotation Sets | Annotations List | Annotations Stack | Co-reference Editor | Text

[Quoted text hidden]
Shelby Schulz <shelby@getnomi.com> Tue, Jul 9, 2013 at 3:26 PM
To: David Leon Rosenthal <rosenthal.dleon@gmail.com>
Cc: Amol Sarva <a@sarva.co>
Hi David,
1pm is great. See you there. If you want some furniture recommendations, I have a few. I also know a company that does furniture rentals if you're interested.
Let me know,
9/27/13 Gmail - 33 5th floor
https://mail.google.com/mail/u/0/?ui=2&ik=6b207d15f2&view=pt&cat=Permathreads&search=cat&th=13fc4be490023a5c 2/4
Shelby
[Quoted text hidden]

Shelby Schulz
getnomi.com
409--692--0005
Amol Sarva <a@sarva.co> Tue, Jul 9, 2013 at 4:19 PM
To: David Leon Rosenthal <rosenthal.dleon@gmail.com>
Hire away
Just ordered 8 chairs
[Quoted text hidden]

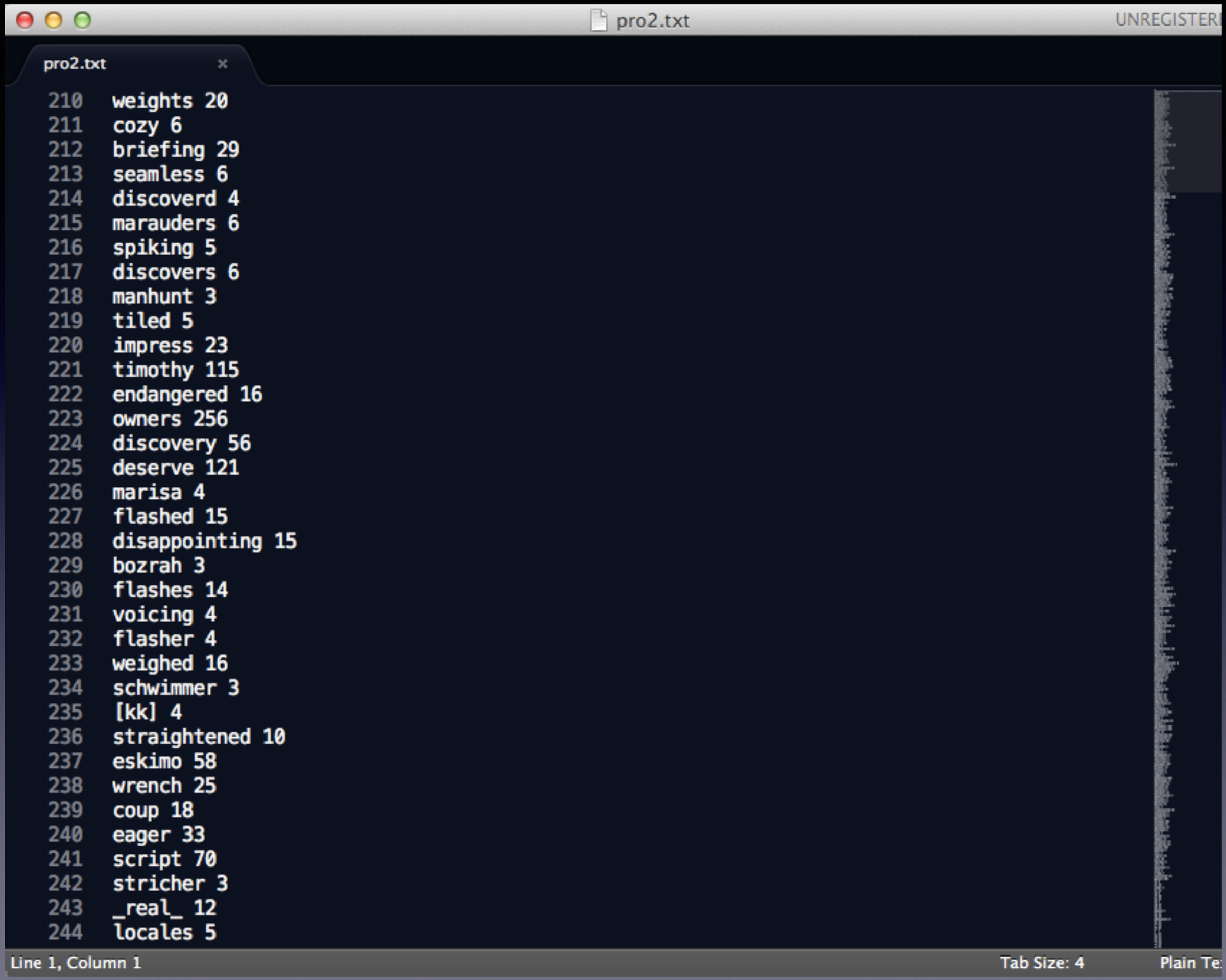
Type	Set	Start	End	Id	Features
UsefulWords		588	593	4921	{string=fired}
UsefulWords		595	597	4922	{string=up}
UsefulWords		610	617	4923	{string=friends}
UsefulWords		621	625	4924	{string=Just}
UsefulWords		627	630	4925	{string=buy}
UsefulWords		632	635	4926	{string=two}
UsefulWords		637	642	4927	{string=power}

422 Annotations (1 selected) Select:

- [ ] Address
- [ ] Date
- [x] Email1
- [ ] FirstPerson
- [ ] Identifier
- [ ] Lookup
- [ ] Money
- [ ] Noise
- [ ] Organization
- [ ] Person
- [x] Quote
- [ ] Sentence
- [ ] SpaceToken
- [ ] Split
- [x] TitleWithLink
- [ ] Token
- [ ] Unknown
- [ ] UrlPre
- [x] UsefulWords
- Original markups

Tf-Idf

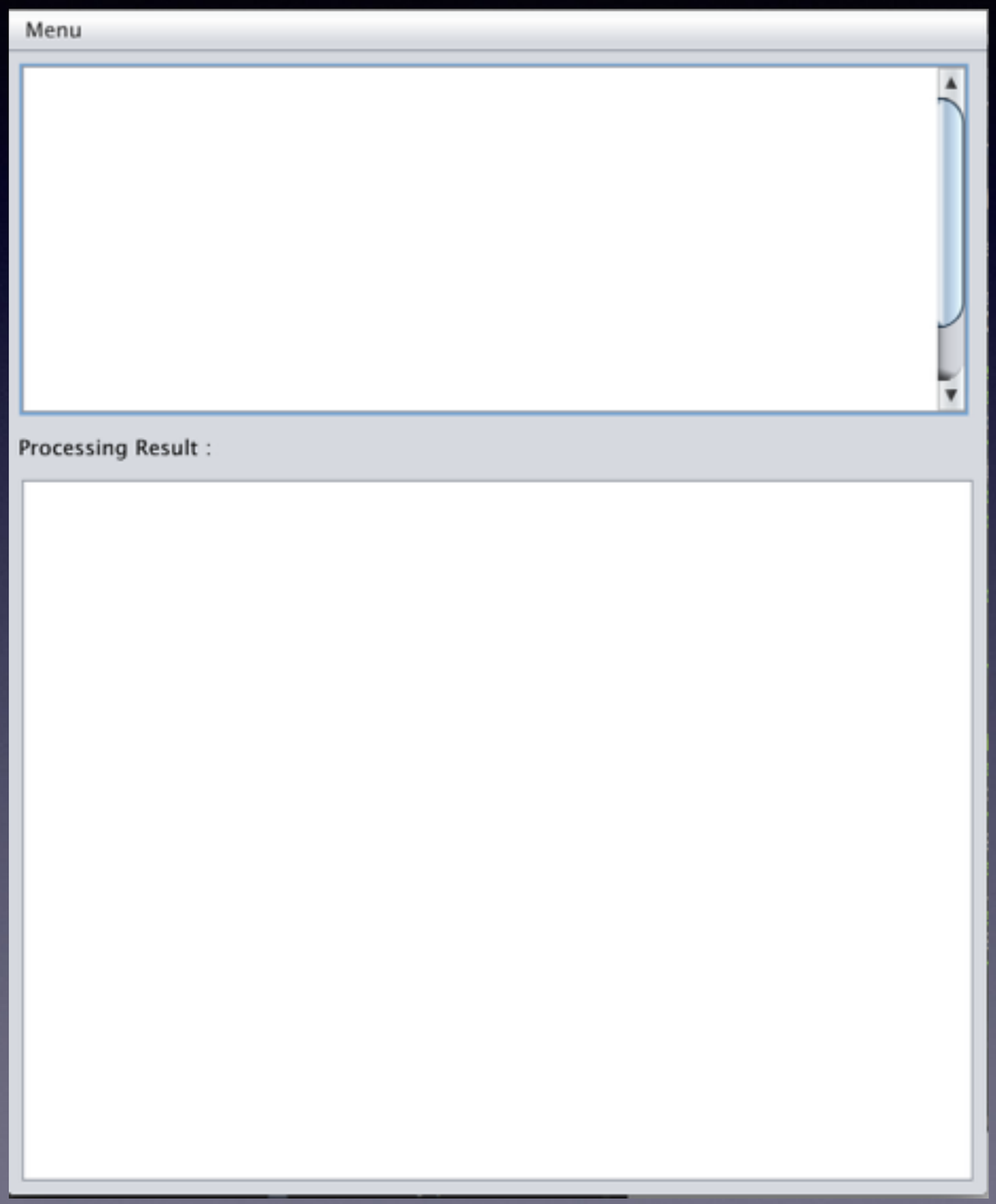
- Term Frequency
  - ✓ Compute the TF for every words in AnnotationSet UsefulWords
  - ✓ Give extra weight to title and first email
- Inverse Document Frequency
  - ✓ Preprocess Newsgroup-18777
  - ✓ Get every words and how many documents it occurs

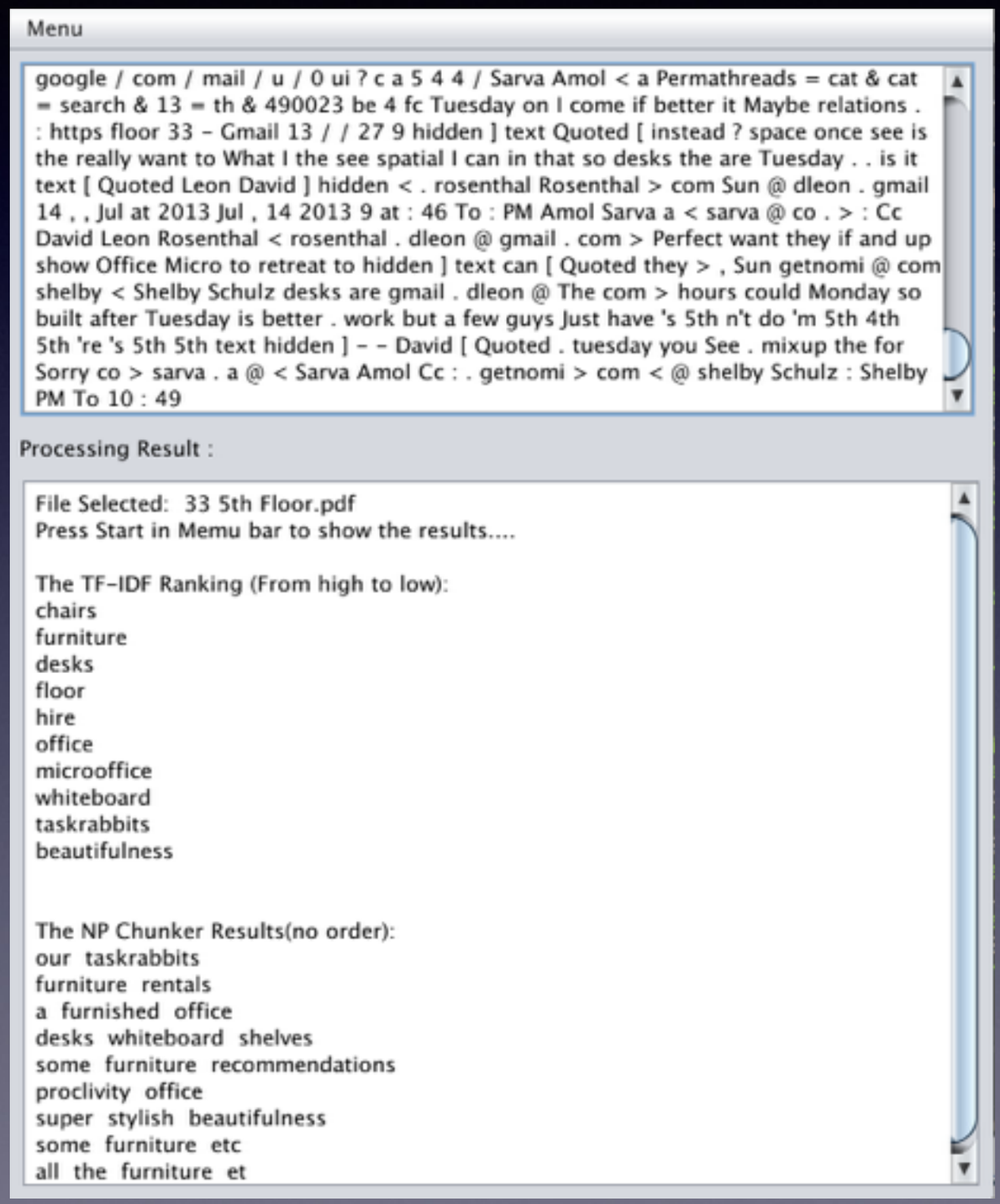

pro2.txt
UNREGISTERE
pro2.txt
210 weights 20
211 cozy 6
212 briefing 29
213 seamless 6
214 discoverd 4
215 marauders 6
216 spiking 5
217 discovers 6
218 manhunt 3
219 tiled 5
220 impress 23
221 timothy 115
222 endangered 16
223 owners 256
224 discovery 56
225 deserve 121
226 marisa 4
227 flashed 15
228 disappointing 15
229 bozrah 3
230 flashes 14
231 voicing 4
232 flasher 4
233 weighed 16
234 schwimmer 3
235 [kk] 4
236 straightened 10
237 eskimo 58
238 wrench 25
239 coup 18
240 eager 33
241 script 70
242 stricher 3
243 _real_ 12
244 locales 5
Line 1, Column 1
Tab Size: 4
Plain Te

Java implementation

- Compute the TF-IDF value for every words
- Save the top 10 TF-IDF value words into a gazetteer
- Run Gate one more time with the new gazetteer
- With Gate plugin Tagger_NP_Chunker, get the noun phase contains the top 10 TF-IDF words

EmailSummar UI





Gate-eclipse connection

- Export the .xgapp file in Gate
- Initialize Gate in Java with the .xgapp
- Initialize the corpus with file loaded in Java GUI
- Get results from Gate and process in Java
- Print out the processing results in GUI

DEMO

Performance

Training set

File Selected: Apt for Video Shoot.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
jana , jpg , room , bedroom , apartment , living , kitchen , shoot , painting , chris ,

The NP Chunker Results(no order):
a bedroom , chris , day shoot , the living room , apartment , bedroom apartment , an apartment , video shoot , the shoot what , your shoot , day shoot

File Selected: ThankYouForOrder.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
delivery , desks , order , fwd , date , serrentino , nomi , total , eiffel , legroom ,

The NP Chunker Results(no order):
the estimated delivery date , the desk order , a guaranteed delivery date , these ikea desks , desks quoted text hidden , more legroom , the actual delivery date , your order , desks quoted text hidden david leon rosenthal rosenthal dleon ,

File Selected: Introduction.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
internship , alzheimer , volunteer , dear , introduction , best , commitments , resume , attached , langone ,

The NP Chunker Results(no order):
an introduction , our volunteer research intern program , your resume , the volunteer program , research or volunteer roles , your commitments and

Tech/Science

File Selected: Google Innovations.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
google , mongodb , knotable , cloud , innovations , amazon , storage , cname , datastore , mongo ,

The NP Chunker Results(no order):
cloud datastore , a knotable subdomain , file storage , cloud storage , google innovations

File Selected: New iOS Design.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
png , vladimirbabic , permissions , groups , ios , design , iphone , group , user , topics ,

The NP Chunker Results(no order):
about the group design , the group permissions settings , dropdown callout permissions menu , iphone profile , ios design , topics profiles , the other topics ,

File Selected: UI speed bugs.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
topics , paging , loading , nga , tasks , threads , column , fetch , bugs , knotes ,

The NP Chunker Results(no order):
off paging , only load knotes , my tasks , loading time , other tasks , all knotes , all the threads , people column , topics column , topics paging , on paging

Daily

File Selected: Birthday.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
cove , urbanspoon , hamish , restaurant , birthday , indian , dad , attadale , peeps , eat ,

The NP Chunker Results(no order):
his birthday , hey peeps , the cove indian restaurant , either restaurant , www urbanspoon

File Selected: Drinks.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
iphone , guys , weekend , mailbox , fridays , village , sent , contact , branch , good ,

The NP Chunker Results(no order):
a good weekend , contact info , contact info , the following weekend , next weekend , looks good , middle branch , some good bars , hey guys , hi guys

File Selected: Cake.pdf
Press Start in Memu bar to show the results....

The TF-IDF Ranking (From high to low):
singtel , blackberry , fabians , midnight , bout , sent , sure , via , raghav , leaving ,

The NP Chunker Results(no order):
our midnight plan

Result

- Technique email: Good
- Daily email: Average

Known Bugs

- Program will run out the memory if processing more than 3-4 documents in a row.

- All the contents were read by the program in text format. Thus cannot get rid of the file suffix such as jpg in email threads containing lots of pictures.

- If an email thread is too short, sometimes the program would fail to find any chunk result.

Shortcomings

- If an unusual trivial term appears in a document many times, it would be easily regarded as a high tf-idf keyword (such as ghz, mpg, etc).

- Cannot generate meaningful short sentences to give the user an intuitive summary of the email thread

- File processing efficiency is still low. Usually costs more than 10 seconds to read in a file. Performance needs to be optimized.

Further Developments

- Expand the noise word list
- Apply more word lists in gazetteer, such as positive word list and negative word list
- Identify the part of speech each word belongs to (subject, object, verb, etc)
- Extract verb-object phases for high tf-idf terms
- Generate short meaningful sentences as the summary

Thank you