RAKE Topic Modeling

Haihao Update

2/10/22

Four tasks

- For each key phrase (and candidate):
  - Where in document (which section)
  - Where in paragraph (beginning/middle/end)

- For key phrase pair co-occurrence (in same paragraph):
  - Order
  - Proximity

Document collection

- Papers on sodium ion batteries (SIB) collated by RM

In [4]:
```
keywords = ['na ion', 'na-ion', 'sodium ion', 'sodium-ion', 'sodium batt']
```

In [5]:
```
paper_set = []

for p in paper_syn:
    if p['title'] is not None and any(k in p['title'].lower() for k in keywords):
        paper_set.append(p)

paper_syn = paper_set
```

Document collection

- Has at least abstract and recipe
- 3150 papers in collection

In [8]:
```
selective_paper_1=[]

for indx1,paper1 in enumerate(paper_syn):
    if paper1['doi'] in screened_list:
        selective_paper_1.append(paper1)
len(selective_paper_1)
```

Out[8]: 3150

- Pickled for export

In [18]:
```
import pickle

with open('SIB_abs_rec.data', 'wb') as fp:
    pickle.dump(abs_rec_f, fp)
```

In [7]:
```
abstract_1=[]
titles=[]
for index, paper in enumerate(paper_syn):
    if paper['title'] is not None:
        if paper['abstract'] is not None:
            abstract_1.append(paper['abstract'])
            recipes=[]
            for para in paper['paragraphs']:
                if para['type']=='recipe':
                    recipes.append(para['text'])
            #print(len(recipes))
            if len(recipes)>1:
                titles.append(paper['doi'])
            else:
                titles.append(paper['doi'])

print(len(titles))
screened_list = list(set(titles))
len(screened_list)
```

```
3150
```

Out[7]: 3150

Document collection

- Three separate lists
  - Abstracts only (`abst`)
  - Recipes only (`rec`)
  - Abstracts + Recipes (`both`)
- 3150 in each
- Same ordering (important)

In [76]:

```
abst=[]
with open('SIB_abs.data', 'rb') as fp:
    abst=pickle.load(fp)
print(len(abst))

rec=[]
with open('SIB_rec.data', 'rb') as fp:
    rec=pickle.load(fp)
print(len(rec))

both=[]
with open('SIB_abs_rec.data', 'rb') as fp:
    both=pickle.load(fp)
print(len(both))
```

```
3150
3150
3150
```

Where in document

- Ran RAKE on set with both abstract and recipe (`both`)
- For each paper:
  - For all candidates and key phrases (top 1/3 of candidates):
    - Regex match in both abstract text only (`abst`) and recipe text only (`rec`)
- If found, track source paragraph type (‘abs’, ‘rec’)
- Can easily include other paragraph types (intro, conclusion)

Where in paragraph

- Regex match (from before) gives locations of all matches
  - Took midpoint of start and end position as location of match
- Location index with paragraph length gives relative position
- Initially tracked all matches
  - Relevant phrases only appear once or twice
  - Noise (e.g. units) appear often, slows down co-occurrence search
- Decided to only track *average* location of all matches

Where in document (candidates)

```
1 t = tracked['cands']['abs'][42]
2 print(len(t))
3 t
```

```
47

[('far received less attention', 749.5, 1194),
 ('organic electrode materials', 802.5, 1194),
 ('low discharge potential', 461.5, 1194),
 ('good reaction reversibility', 640.5, 1194),
 ('next generation green', 76.5, 1194),
 ('quinone electrode materials', 706.5, 1194),
 ('high resource availability', 672.0, 1194),
 ('high theoretical capacity', 612.5, 1194),
 ('multifaceted modification approaches', 1092.0, 1194),
 ('poor electronic conductivity', 430.0, 1194),
 ('sodium ion batteries', 122.0, 1194),
 ('electrode materials', 511.5, 1194),
 ('redox stability', 229.5, 1194),
 ('electronic conductivity', 730.0, 1194),
 ('postive materials', 485.5, 1194),
 ('sustainable lithium', 101.5, 1194),
 ('low cost', 164.0, 1194),
 ('discharge plateaus', 1054.0, 1194)
```

```
1 t = tracked['cands']['rec'][42]
2 print(len(t))
3 t
```

```
148

[('highly conductive carbon matrices facilitates', 178.5, 4256),
 ('green humate lithium electrode material', 2402.5, 4256),
 ('emodin active material loading delivered', 1716.0, 4256),
 ('strong p p interactions', 2124.5, 4256),
 ('nanocomposite electrode material demonstrated', 1069.5, 4256),
 ('showing poor cyclic stability', 2859.5, 4256),
 ('three electrode materials exhibited', 3937.5, 4256),
 ('single wall carbon nanotube', 1499.5, 4256),
 ('indicating good cycling performance', 1315.5, 4256),
 ('excellent electronic conducting agent', 3094.5, 4256),
 ('high active materials load', 802.0, 4256),
 ('low average discharge potential', 2506.5, 4256),
 ('discharge capacity decreasing sharply', 2792.5, 4256),
 ('high initial discharge capacity', 1110.5, 4256),
 ('h g 1 owing', 666.5, 4256),
 ('full battery demonstrated', 2636.5, 4256),
 ('three electrode materials', 4014.5, 4256),
 ('good cycling stability', 2473.0, 4256)
```

Where in document (keywords)

```
1 t = tracked['kws']['abs'][42]
2 print(len(t))
3 t
```

```
15

[('far received less attention', 749.5, 1194),
 ('organic electrode materials', 802.5, 1194),
 ('low discharge potential', 461.5, 1194),
 ('good reaction reversibility', 640.5, 1194),
 ('next generation green', 76.5, 1194),
 ('quinone electrode materials', 706.5, 1194),
 ('high resource availability', 672.0, 1194),
 ('high theoretical capacity', 612.5, 1194),
 ('multifaceted modification approaches', 1092.0, 1194),
 ('poor electronic conductivity', 430.0, 1194),
 ('sodium ion batteries', 122.0, 1194),
 ('electrode materials', 511.5, 1194),
 ('redox stability', 229.5, 1194),
 ('electronic conductivity', 730.0, 1194),
 ('postive materials', 485.5, 1194)]
```

```
1 t = tracked['kws']['rec'][42]
2 print(len(t))
3 t
```

```
40

[('highly conductive carbon matrices facilitates', 178.5, 4256),
 ('green humate lithium electrode material', 2402.5, 4256),
 ('emodin active material loading delivered', 1716.0, 4256),
 ('strong p p interactions', 2124.5, 4256),
 ('nanocomposite electrode material demonstrated', 1069.5, 4256),
 ('showing poor cyclic stability', 2859.5, 4256),
 ('three electrode materials exhibited', 3937.5, 4256),
 ('single wall carbon nanotube', 1499.5, 4256),
 ('indicating good cycling performance', 1315.5, 4256),
 ('excellent electronic conducting agent', 3094.5, 4256),
 ('high active materials load', 802.0, 4256),
 ('low average discharge potential', 2506.5, 4256),
 ('discharge capacity decreasing sharply', 2792.5, 4256),
 ('high initial discharge capacity', 1110.5, 4256),
 ('h g 1 owing', 666.5, 4256),
```

Co-occurrences

- Completely new way of searching, drastically faster

- Old:
  - Rank all keywords in collection (by `tid`), choose top n = 100/1000/etc.
  - Search for all O(n^2) possible pairs in every paper in collection (SLOW!)

- New:
  - *Create* all O(n^2) possible pairs with tracked locations from one paper ($n < 100$)
  - Rank pairs across collection *afterwards* (by frequency is easiest, or scores)

Order and proximity

- Already have locations from regex match earlier
- Iterate over all $n(n-1)/2$ possible combinations, retaining order
- **Proximity** is simply difference in two (average) match locations
- **Order** is simply sign of proximity (convention: second – first)

Co-occurrences in abstract

```
1  c = co_occs['abs'][42]
2  print(len(c))
3  pp.pprint(c,width=120)
```

```
105
[('far received less attention', 'organic electrode materials', 53.0, 1194),
 ('far received less attention', 'low discharge potential', -288.0, 1194),
 ('far received less attention', 'good reaction reversibility', -109.0, 1194),
 ('far received less attention', 'next generation green', -673.0, 1194),
 ('far received less attention', 'quinone electrode materials', -43.0, 1194),
 ('far received less attention', 'high resource availability', -77.5, 1194),
 ('far received less attention', 'high theoretical capacity', -137.0, 1194),
 ('far received less attention', 'multifaceted modification approaches', 342.5, 1194),
 ('far received less attention', 'poor electronic conductivity', -319.5, 1194),
 ('far received less attention', 'sodium ion batteries', -627.5, 1194),
 ('far received less attention', 'electrode materials', -238.0, 1194),
 ('far received less attention', 'redox stability', -520.0, 1194),
 ('far received less attention', 'electronic conductivity', -19.5, 1194),
 ('far received less attention', 'postive materials', -264.0, 1194),
 ('organic electrode materials', 'low discharge potential', -341.0, 1194),
 ('organic electrode materials', 'good reaction reversibility', -162.0, 1194),
 ('organic electrode materials', 'next generation green', -726.0, 1194),
 ('organic electrode materials', 'quinone electrode materials', -96.0, 1194),
```

Co-occurrences in recipe

```
1 c = co_occs['rec'][42]
2 print(len(c))
3 pp.pprint(c,width=120)
```

```
780
[('highly conductive carbon matrices facilitates', 'green humate lithium electrode material', 2224.0, 4256),
 ('highly conductive carbon matrices facilitates', 'emodin active material loading delivered', 1537.5, 4256),
 ('highly conductive carbon matrices facilitates', 'strong p p interactions', 1946.0, 4256),
 ('highly conductive carbon matrices facilitates', 'nanocomposite electrode material demonstrated', 891.0, 4256),
 ('highly conductive carbon matrices facilitates', 'showing poor cyclic stability', 2681.0, 4256),
 ('highly conductive carbon matrices facilitates', 'three electrode materials exhibited', 3759.0, 4256),
 ('highly conductive carbon matrices facilitates', 'single wall carbon nanotube', 1321.0, 4256),
 ('highly conductive carbon matrices facilitates', 'indicating good cycling performance', 1137.0, 4256),
 ('highly conductive carbon matrices facilitates', 'excellent electronic conducting agent', 2916.0, 4256),
 ('highly conductive carbon matrices facilitates', 'high active materials load', 623.5, 4256),
 ('highly conductive carbon matrices facilitates', 'low average discharge potential', 2328.0, 4256),
 ('highly conductive carbon matrices facilitates', 'discharge capacity decreasing sharply', 2614.0, 4256),
 ('highly conductive carbon matrices facilitates', 'high initial discharge capacity', 932.0, 4256),
 ('highly conductive carbon matrices facilitates', 'h g 1 owing', 488.0, 4256),
 ('highly conductive carbon matrices facilitates', 'full battery demonstrated', 2458.0, 4256),
 ('highly conductive carbon matrices facilitates', 'three electrode materials', 3836.0, 4256),
 ('highly conductive carbon matrices facilitates', 'good cycling stability', 2294.5, 4256),
 ('highly conductive carbon matrices facilitates', 'high initial capacity', 353.0, 4256),
```

Next steps (how to prioritize)

- Include other paragraph types, recalculate counts and scores
  - For both individual key phrases and pairs
- See which phrases most unique to abstracts, recipes, etc.
  - As was done for alloy series (via `tid` score)
- Rank co-occurrence pairs across entire collection
  - Look for patterns in order and proximity in most common pairs