SOCRATex
Staged Optimization of Curation, Regularization, and Annotation of clinical Text

Jimyung Park¹, Seng Chan You M.D. M.S.², Jin Roh M.D. Ph.D.⁴, Rae Woong Park M.D. Ph.D.¹,²,³

¹ Dept. of Biomedical Sciences, Ajou University Graduate School of Medicine, Yeongtong-gu, Suwon, 16499
² Dept. of Biomedical Informatics, Ajou University School of Medicine, Yeongtong-gu, Suwon, 16499
³ FEEDER-NET(Federated E-Health Big Data for Evidence Renovation Network)
⁴ Dept. of Pathology, Ajou University Hospital, Yeongtong-gu, Suwon, 16499
Background

• State-of-the-art (SOTA) methods made great stories on Natural Language Processing (NLP) tasks

• Yet, the SOTA methods usually require massive amounts of labeled-data to learn

• SOCRATex is a Natural Language Processing (NLP) system which helps users to understand, annotate, and retrieve their text documents
System Architecture of SOCRATex

Phase I
- removeNonEnglish
- removeNumbers
- stripWhitespace
- removePunctuation
- user-defined dictionary

Preprocessing

Phase II
- Text Clustering
- Define Data Structure
- colon pathology
  - anatomic site
  - histology
  - biomarker test

Phase III
- Clinical narrative document
  - gross, received in formalin. biopsy result, colon, sigmoid, adenocarcinoma, well differentiated. ** additional note) kras mutation analysis.

{ "colon pathology" : {
  "anatomic site" : "colon, sigmoid",
  "histology" : "adenocarcinoma",
  "biomarker test" : "kras"
} }
Latent Dirichlet Allocation

- Latent Dirichlet Allocation (LDA) is a statistical topic model method which captures latent topics in documents.

I. LDA is an unsupervised method which can reduce time and cost of reviewing the reports.

II. LDA results can give an insight how to annotate and re-construct their text data.
Elastic Stack

- **Elasticsearch** is an open-source indexing framework for searching, which can accept tree-based documents such as JSON documents.
- **Kibana** helps to visualize indexed data stored in Elasticsearch.
Data Source

- Ajou University Medical Center
- ICD-10th C18-20 diagnosed patients from 2014-2017 were included
- 1,989 pathology reports on colorectal cancer of 1,929 patients were included
Results of LDA

margin, resection, mass, lymph, invasion, node, regional, xcm, metastasis, distal, apart, len, pericolic, identified, proximal, carcinoma, circumference, nodes, fresh, some, free, illdefined, state, cut, instability, test, msimicrosatellite, bat, invades, iple
Results of LDA

- **kras**, mutation, analysis, dna, realtime, clamping, pcr, codon, comments, antiegrf, rapy, msi, using, genomic, isolated, mediated, paraffinembedded, target, cetuximab, panitumumab, marker, pnamediated, materials, erlotinib, gefitinib, kinase, tyrosine, inhibitor, pna, additional
## Results of LDA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Expertise Annotation</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic1</td>
<td>Malignant, biopsy</td>
<td>biopsy, all, consists, xcm, embedded, mucosal, received, measuring, diagnosis, sections, tissue, pieces, labelled, gross, biopsied, adenocarcinoma, cancer, differentiated, colon, moderately, rectal, verge, rectum, anal, four, sigmoid, endoscopic, largest, five, one</td>
</tr>
<tr>
<td>Topic2</td>
<td>Benign, biopsy</td>
<td>anal, verge, colon, one, tubular, adenoma, low, grade, dysplasia, biopsy, transverse, polypectomy, containers, each, ascending, identified, consists, two, polyp, largest, sigmoid, descending, polypoid, hyperplastic, mucosal, proximal, endoscopic, polyps, xcm, three</td>
</tr>
<tr>
<td>Topic3</td>
<td>Lymph node invasion, surgery</td>
<td>margin, resection, mass, lymph, invasion, node, regional, xcm, metastasis, distal, apart, len, pericolic, identified, proximal, carcinoma, circumference, nodes, fresh, some, free, illdefined, state, cut, instability, test, msimicrosatellite, bat, invades, iple</td>
</tr>
<tr>
<td>Topic4</td>
<td>Cancer, surgery</td>
<td>invasion, adenoma, resection, margin, submitted, consu, ation, hampe, grade, histopathologic, stained, size, adenocarcinoma, dysplasia, high, tumor, tublovillous, type, depth, low, biopsy, gross, well, tubular, labelled, differentiated, polypectomy, colon, endoscopic, whitish</td>
</tr>
<tr>
<td>Topic5</td>
<td>Gene mutation analysis</td>
<td>kras, mutation, analysis, dna, realtime, clamping, pcr, codon, comments, antigfr, rapy, msi, using, genomic, isolated, mediated, paraffinembedded, target, cetuximab, panitumumab, marker, pna mediated, materials, erlotinib, gefitinib, kinase, tyrosine, inhibitor, pna, additional</td>
</tr>
</tbody>
</table>
Defining JSON structure based on LDA

**Results**

**Topic 1**
- Rectum, endoscopic biopsy: tubulovillous adenoma with high grade dysplasia and adenocarcinoma

**Topic 2**
- Colon, proximal transverse, polypectomy: Hyperplastic polyp. Tubular adenoma with low grade dysplasia and clear resection margin.

**Topic 3**
- Colon, radical sigmoid colectomy: Adenocarcinoma
- Regional lymph node metastasis: no metastasis in all 17 regional lymph node nodes (pN0) (pericolic 0/17) Lymphatic invasion: not identified

**Topic 4**
- Colon and rectum, Hartman’s operation: Adenocarcinoma, moderately differentiated.
  1. Location: rectum 2. Gross type: ulcerofungating 3. Size: 7.5x6cm 4. Invased perirectal adipose tissue

**Topic 5**
- Additional report, Kras Mutation Analysis Report; Kras mutation is not detected by PNA mediated real-time PCR clamping method.
  Materials: Genomic DNA isolated from paraffin-embedded tissue
Clinical Analysis using the Annotated JSON Documents

- Using two keys of JSON structure, **TNM stage** was easily extracted, and **5-year survival analysis** was conducted
  - depth of invasion
  - metastasis lymph node
- This research shows **SOCRA Tex** can actually be combined with **OMOP CDM** and help to conduct **clinical analysis**
- **SOCRA Tex** is available at [https://github.com/ABMI/SOCRA Tex](https://github.com/ABMI/SOCRA Tex)

### Numbers at risk

<table>
<thead>
<tr>
<th></th>
<th>350</th>
<th>242</th>
<th>180</th>
<th>118</th>
<th>73</th>
<th>34</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I, II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage III</td>
<td>107</td>
<td>84</td>
<td>59</td>
<td>44</td>
<td>30</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

Thank You