Semantic Search Engine for Herbal Medicine

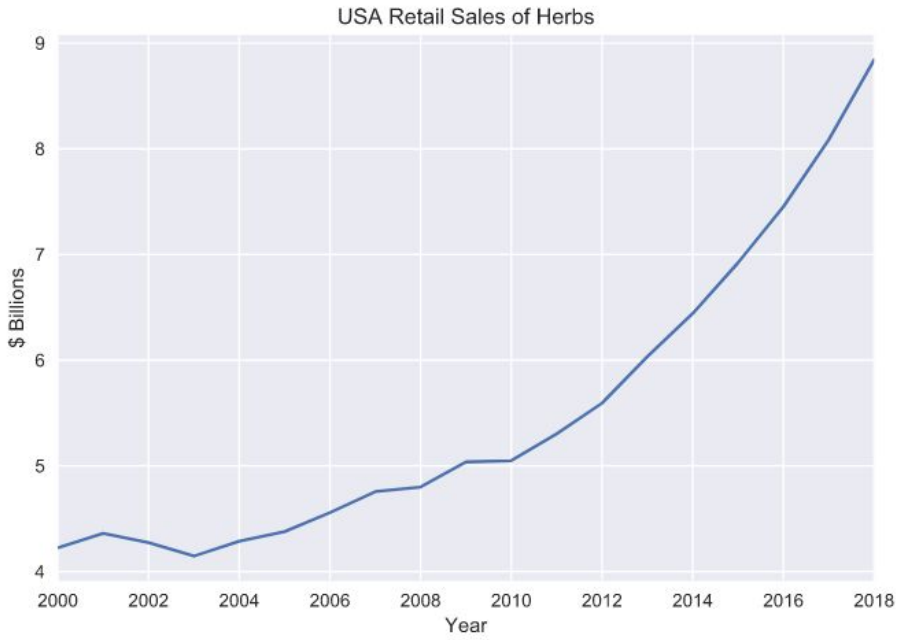
Team Herbert: Emmy, Gurdit, Ian, Karl

CONTENTS

BUSINESS OVERVIEW

HERBAL AND ALTERNATIVE MEDICINE

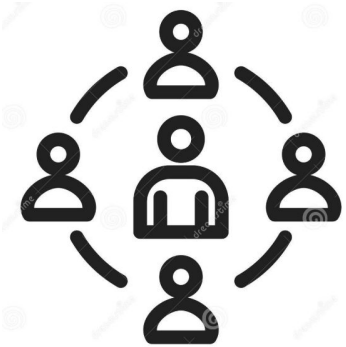
- 1 billion health queries daily on Google, accounting for 7% of traffic
- Alternative medicine accounts for about 1/3 of health search topics.
- Herbal pharmaceuticals sales accounted for $8.8 billion in 2018 domestically and growth has been accelerating over the past decade.



- Where to look?
- Who to trust?
- How to align information on different websites?

Aggregate and distill the most reliable herbal medicine information into a **curated** single report view

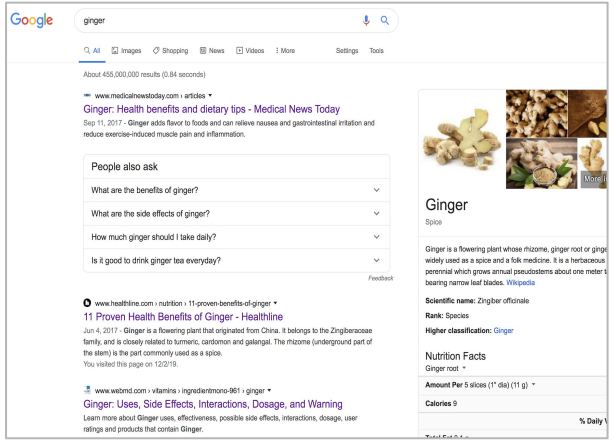
---



User who is curious about **traditional chinese medicine** and finding credible information on its **benefits** and **risks**

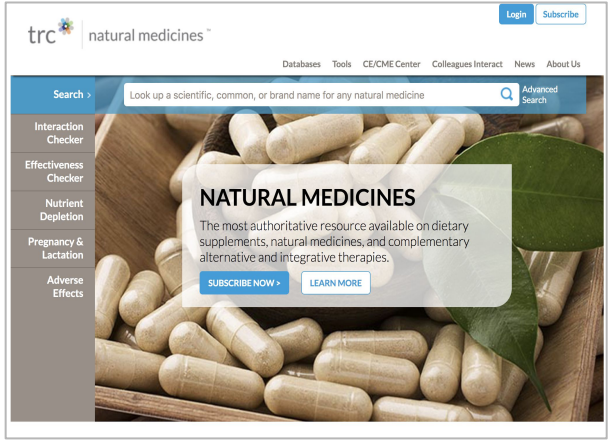
MINIMUM VIABLE PRODUCT

CURRENT SOLUTIONS



GENERAL PURPOSE SEARCH ENGINE

HERBAL MEDICINE WEBSITE



PROFESSIONAL MEDICAL WEBSITE

Too much data

Potentially low quality data

Overly technical data

OUR DIFFERENTIATIONS

VERTICAL SEARCH ENGINE	Herbert focuses on the search for semantic relationship between herbs and conditions
SEMANTIC RELATIONSHIP	Herbert reliability extracts, identifies and develops semantic understandings of 3 main entities for each herbal medicine: herb, conditions and interactions with medications
MULTIPLE DATA SOURCES	Herbert aggregates herbal medicine information from multiple data sources and prioritizes based on the number of data sources that reference them
LINKS TO ORIGINAL DATA	Herbert provides links back to the original data source whenever possible for user to get more information

DATA SOURCES

Scientific Journal of peer-reviewed biomedical literature

Chinese Herbal Medicine Database

MedlinePlus is an online health information service of the National Library of Medicine (NLM)

Integrative database of traditional Chinese medicine (TCM) focused on symptom mapping

Online encyclopedia where most general public use to find health information

MINIMUM VIABLE PRODUCT

2 PRIMARY USE CASES

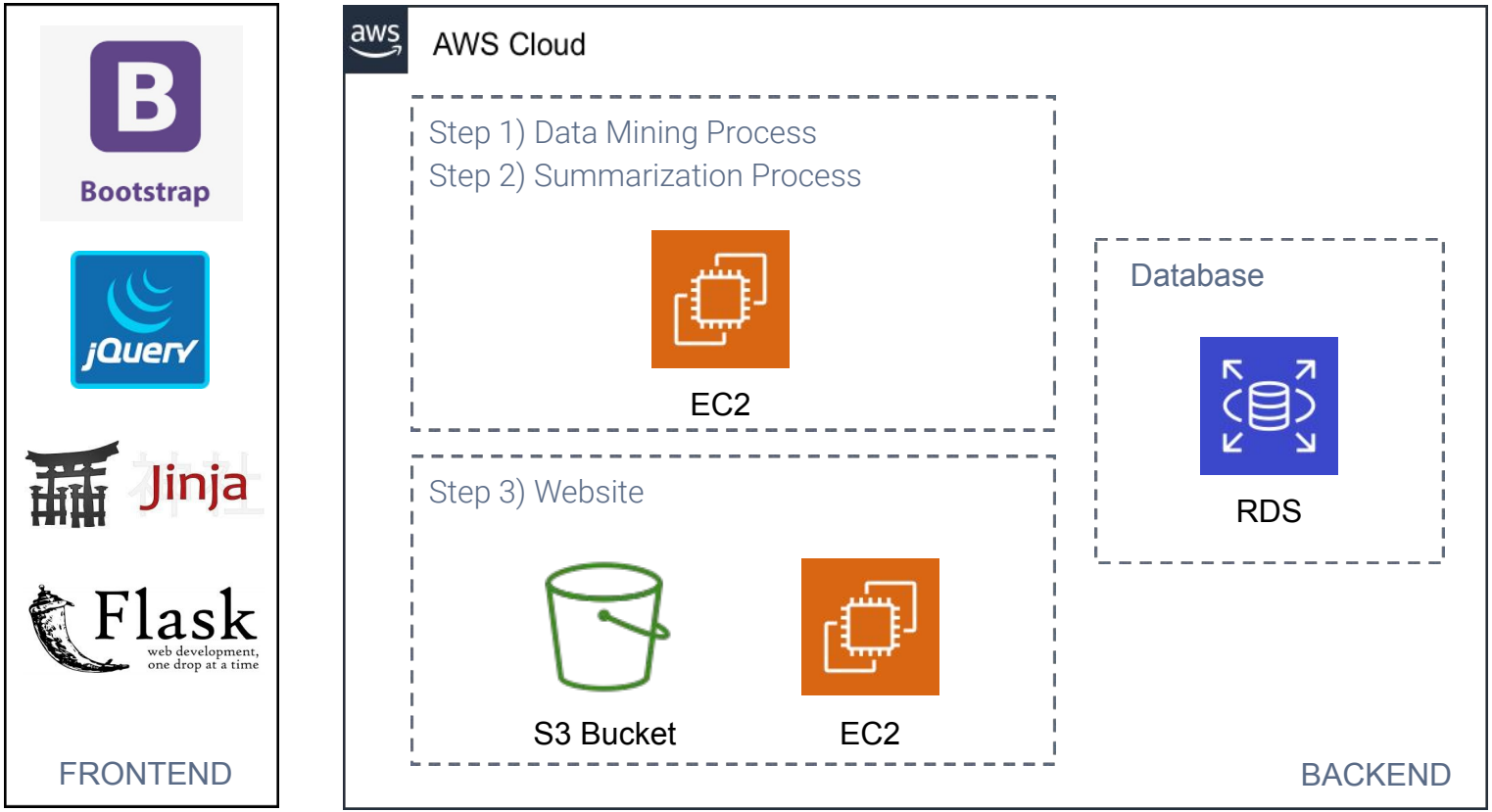
Search for information about a herb

Search for herbs for a particular condition

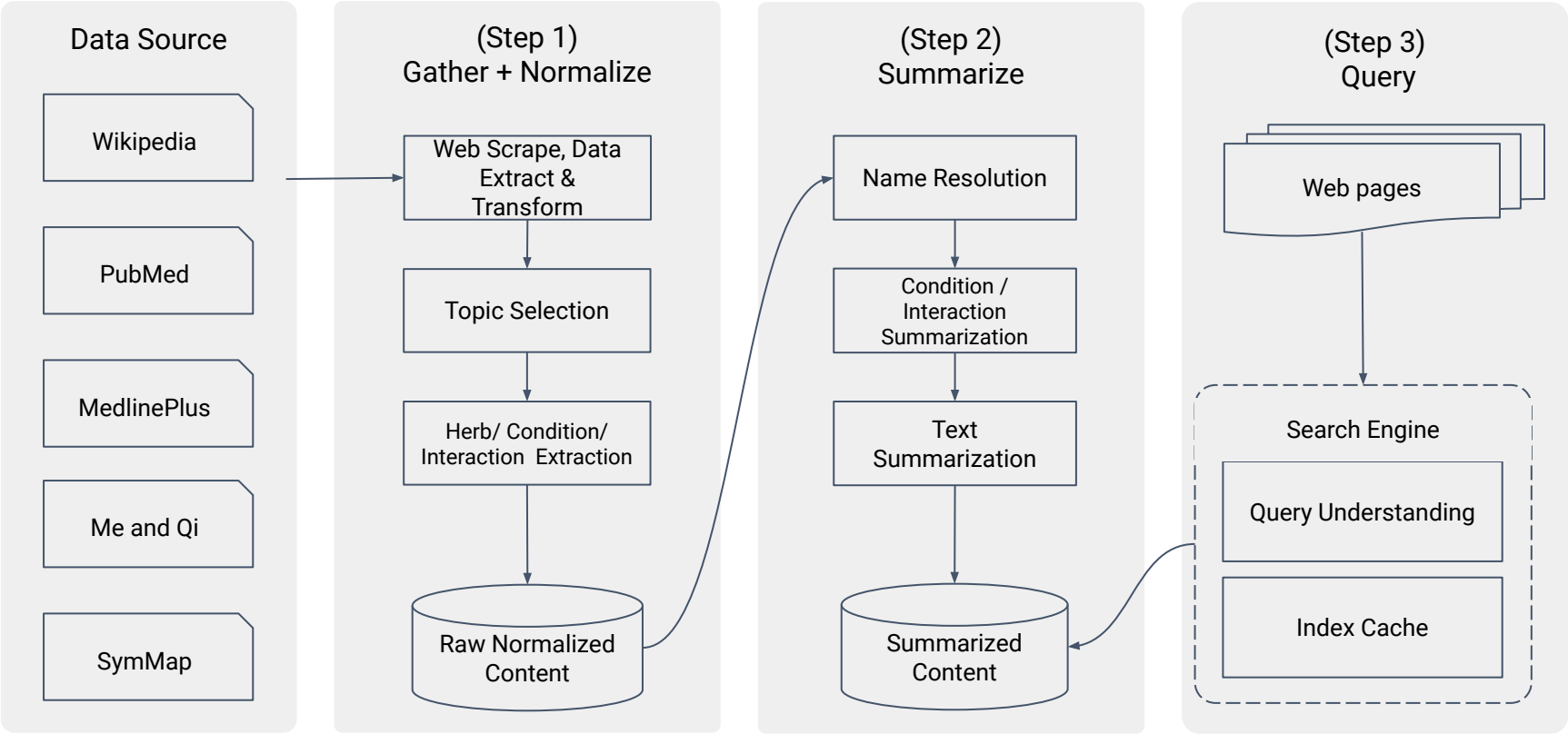
PRODUCT DEMO

DATA SCIENCE/ TECHNICAL APPROACH

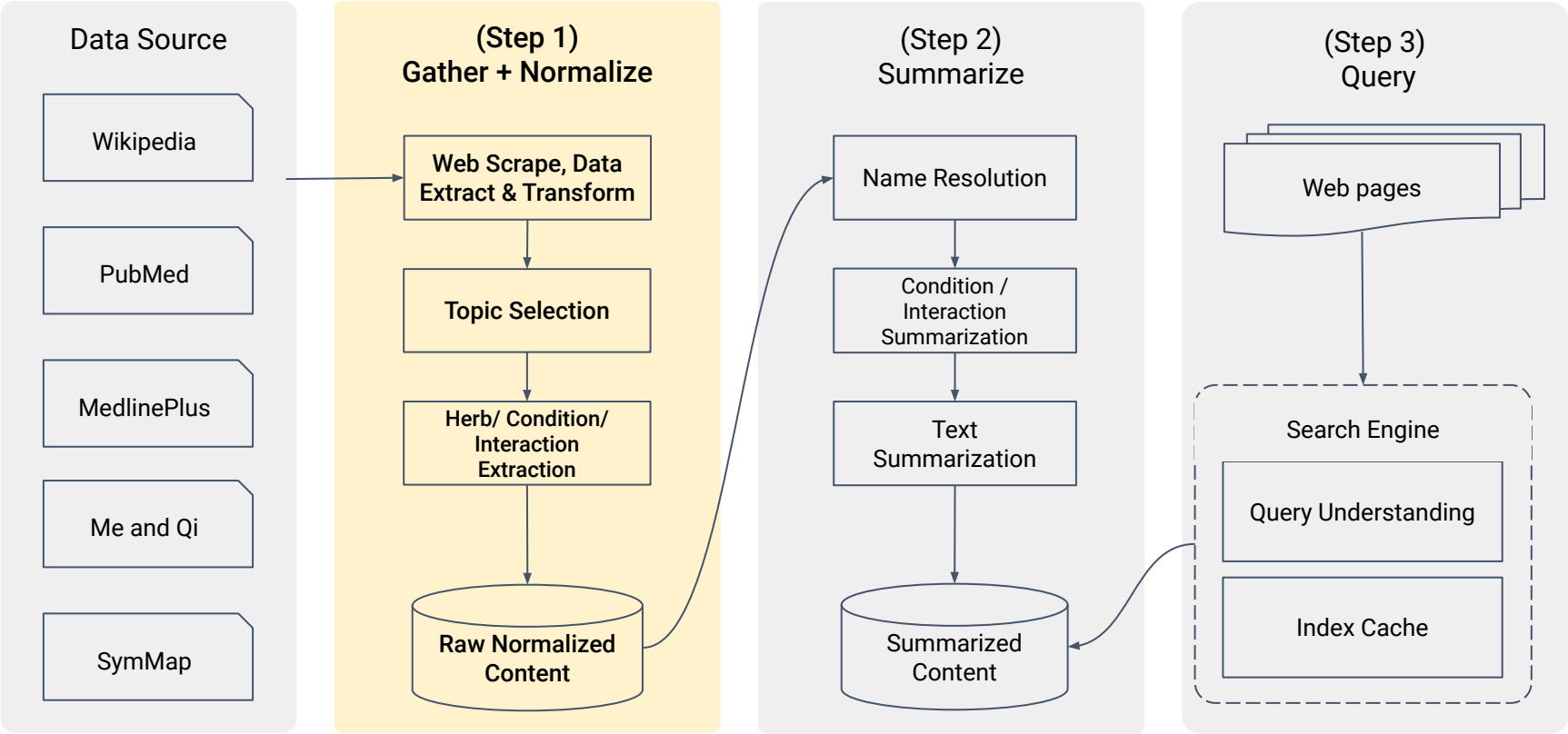
HERBERT AWS ARCHITECTURE



HERBERT OVERALL ARCHITECTURE



(STEP 1) GATHER + NORMALIZE



Step 1: Distill source data to relevant bullet points about an herb

en.wikipedia.org/wiki/Ginger

Bookmarks | Cumberland Consul... | Mail - Ian Bettinger... | OpenAir : Home : C... | iNestIT | PCV | PCV Data Tracker | Gen

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

In other projects

Wikimedia Commons
Wikispecies

Print/export

Create a book
Download as PDF
Printable version

Languages

Deutsch
Español
Français

transported with them throughout the Indo-Pacific during the Austronesian expansion (c. 5,000 BP), reaching as
Ginger is one of the first spices to have been exported from Asia, arriving in Europe with the spice trade, and wa
ancient Greeks and Romans.[4] The distantly related dicots in the genus *Asarum* are commonly called wild ginge
their similar taste.

Contents [hide]

Etymology

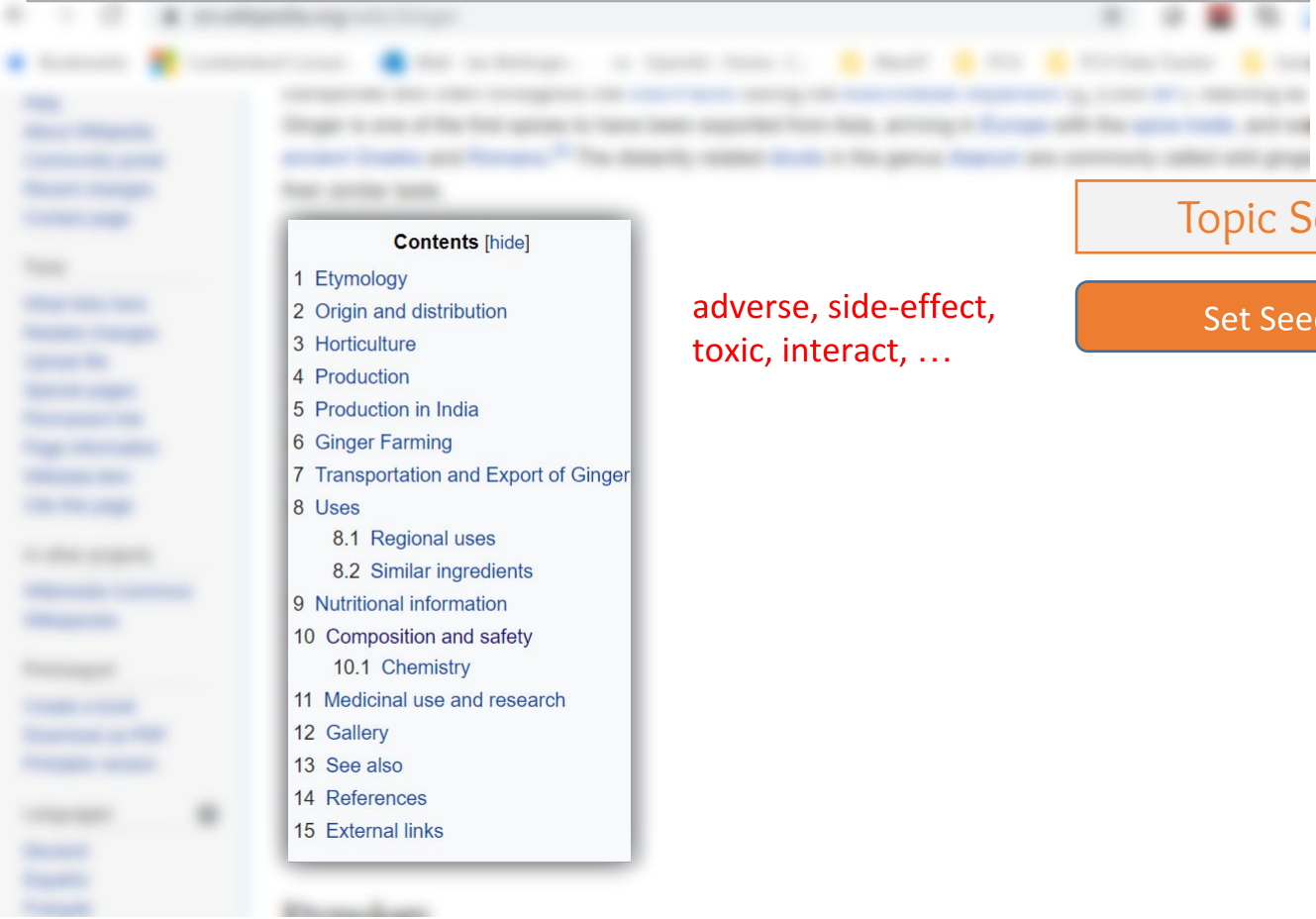
Topic Selection

Step 1: Distill source data to relevant bullet points about an herb

Topic Selection

Contents [hide]

Step 1: Distill source data to relevant bullet points about an herb



adverse, side-effect, toxic, interact, …

Topic Selection

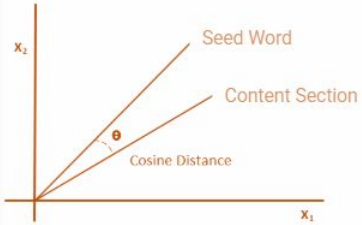
Set Seed Words

Step 1: Distill source data to relevant bullet points about an herb

Contents [hide]

adverse, side-effect, toxic, interact, …

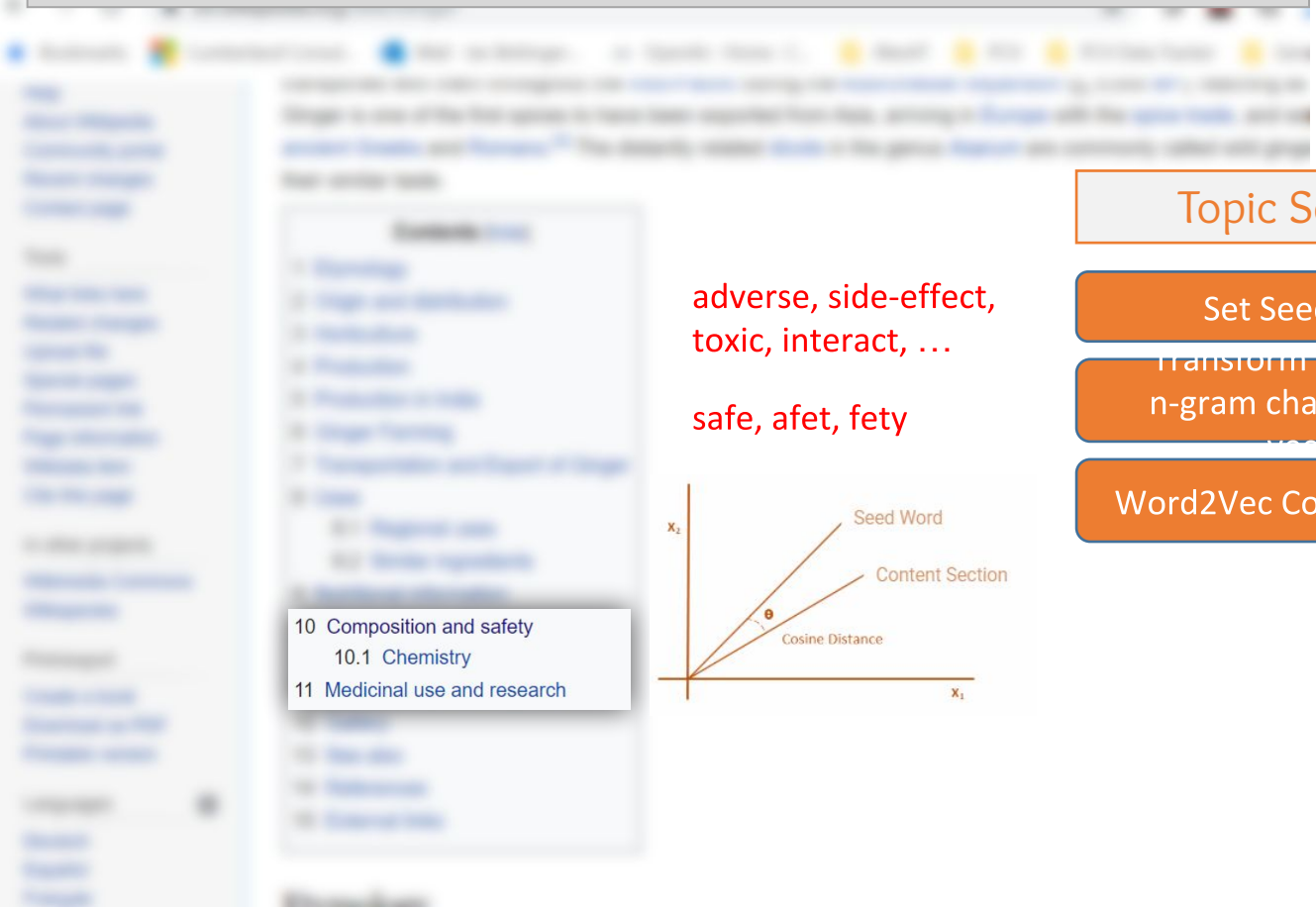
safe, afet, fety



Topic Selection

- Set Seed Words
- Transform Contents to n-gram character word vectors
- Word2Vec Cosine Similarity

Step 1: Distill source data to relevant bullet points about an herb



adverse, side-effect, toxic, interact, …

safe, afet, fety

Topic Selection

- Set Seed Words
- Transform Contents to n-gram character word vectors
- Word2Vec Cosine Similarity

Step 1: Distill source data to relevant bullet points about an herb

Composition and safety

If consumed in reasonable quantities, ginger has few negative side effects.[37] It is on the FDA's "generally recognized as safe" list,[38] though it does interact with some medications, including the anticoagulant drug warfarin[39] and the cardiovascular drug nifedipine.[2]

Chemistry

The characteristic fragrance and flavor of ginger result from volatile oils primarily consisting of zingerone, shogaols, and gingerols with [6]-ginge decanone) as the major pungent compound.[40] Zingerone is produced pungency and a spicy-sweet aroma.[40] Shagoals are more pungent an raw ginger, but is formed from gingerols during heating, storage or via acidity.[40]

Fresh ginger also contains an enzyme zingibain which is a cysteine protease and has similar properties to rennet.

Content Parsing

Find Medical Sentence

Medicinal use and research

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, although its safety has not been established.[41][44] Ginger is not effective for treating dysmenorrhea,[45] and there is insufficient evidence for it having analgesic properties due to the lack of well conducted trials. Available data provides weak evidence for its anti-inflammatory role and it may reduce the subjective experience of pain in osteoarthritis.[46]

Allergic reactions to ginger generally result in a rash.[2] Although generally recognized as safe, ginger can cause heartburn and other side effects, particularly if taken in powdered form.[2] It may adversely affect individuals with gallstones and may interfere with the effects of anticoagulants, such as warfarin or aspirin.[2]

Pantothenic acid (B_5)	0.203 mg	4%
Vitamin B_6	0.16 mg	12%
Folate (B_9)	11 µg	3%
Vitamin C	5 mg	6%
Vitamin E	0.26 mg	2%
Minerals	**Quantity**	**%DV†**
Calcium	16 mg	2%
Iron	0.6 mg	5%
Magnesium	43 mg	12%
Manganese	0.229 mg	11%
Phosphorus	34 mg	5%
Potassium	415 mg	9%
Sodium	13 mg	1%
Zinc	0.34 mg	4%
Other constituents	**Quantity**	
Water	79 g	

Full link to USDA Database entry

Units
µg = micrograms • mg = milligrams
IU = International units

†Percentages are roughly approximated using US recommendations for adults.
Source: USDA Nutrient Database

Step 1: Distill source data to relevant bullet points about an herb

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, although its safety has

Content Parsing

Find Medical Sentence

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, although its safety has

It may adversely affect individuals with gallstones and may interfere with the effects of anticoagulants, such as warfarin or aspirin.[2]

Step 1: Distill source data to relevant bullet points about an herb

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, althoug

Content Parsing

Find Medical Sentence

Relation Extraction

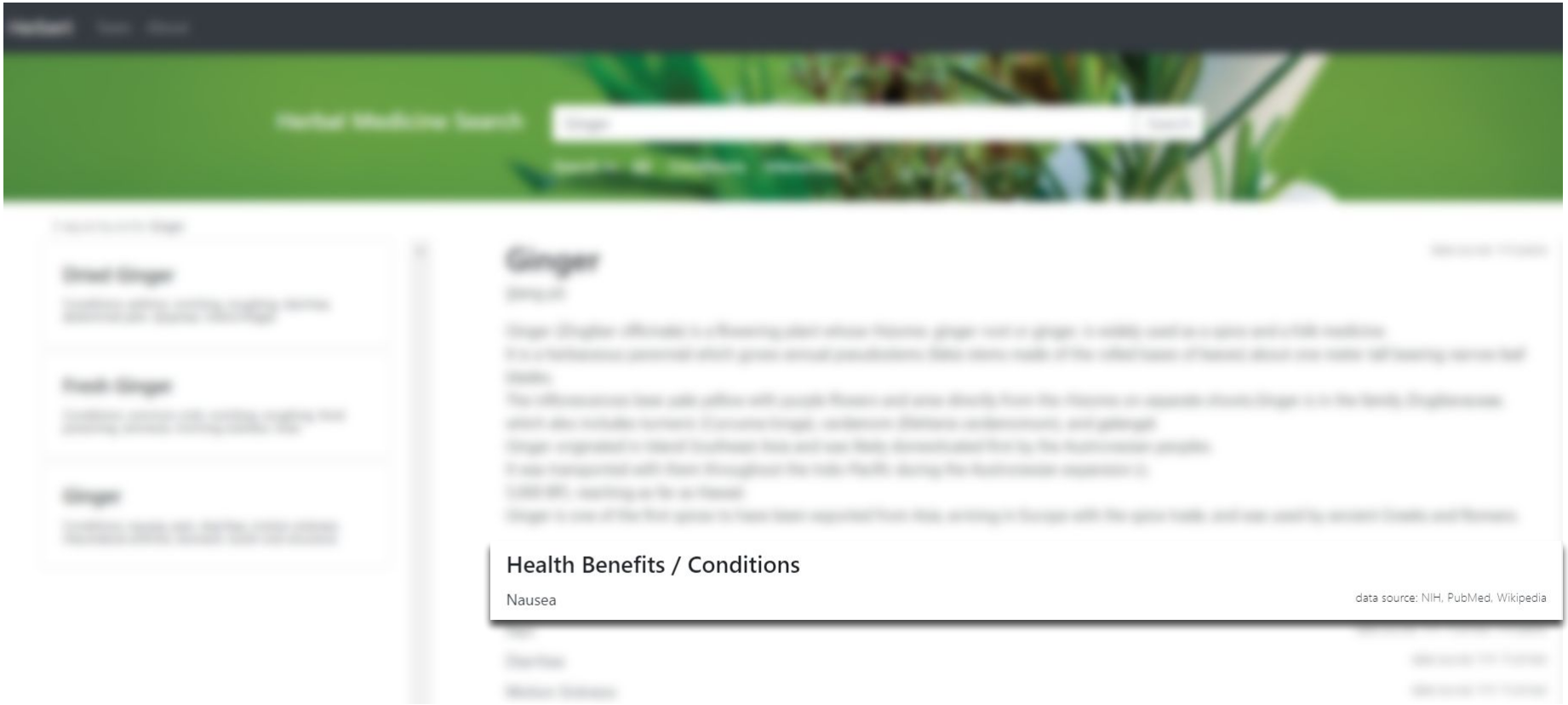
Step 1: Distill source data to relevant bullet points about an herb

Evidence that ginger helps alleviate nausea and vomiting resulting from chemotherapy or pregnancy is inconsistent.[2][41][42][43] There is no clear evidence of harm from taking ginger during pregnancy, althoug

Ginger **alleviates** **nausea**

Content Parsing

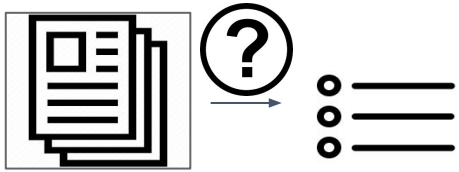
- Find Medical Sentence
- Relation Extraction
- Normalize the data


Herbal Medicine Search
Ginger
Dried Ginger
Fresh Ginger
Ginger
Health Benefits / Conditions
Nausea
data source: NIH, PubMed, Wikipedia

(STEP 1) CHALLENGES AND LEARNING

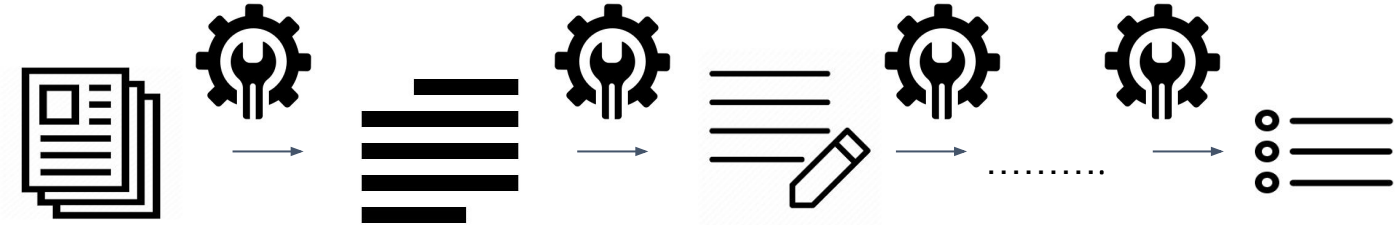
Challenges:

- Grabbing the relevant text
- no reliable single-step solutions
- working without “labeled” data
- working cross-domains (medical + cultural aspects)

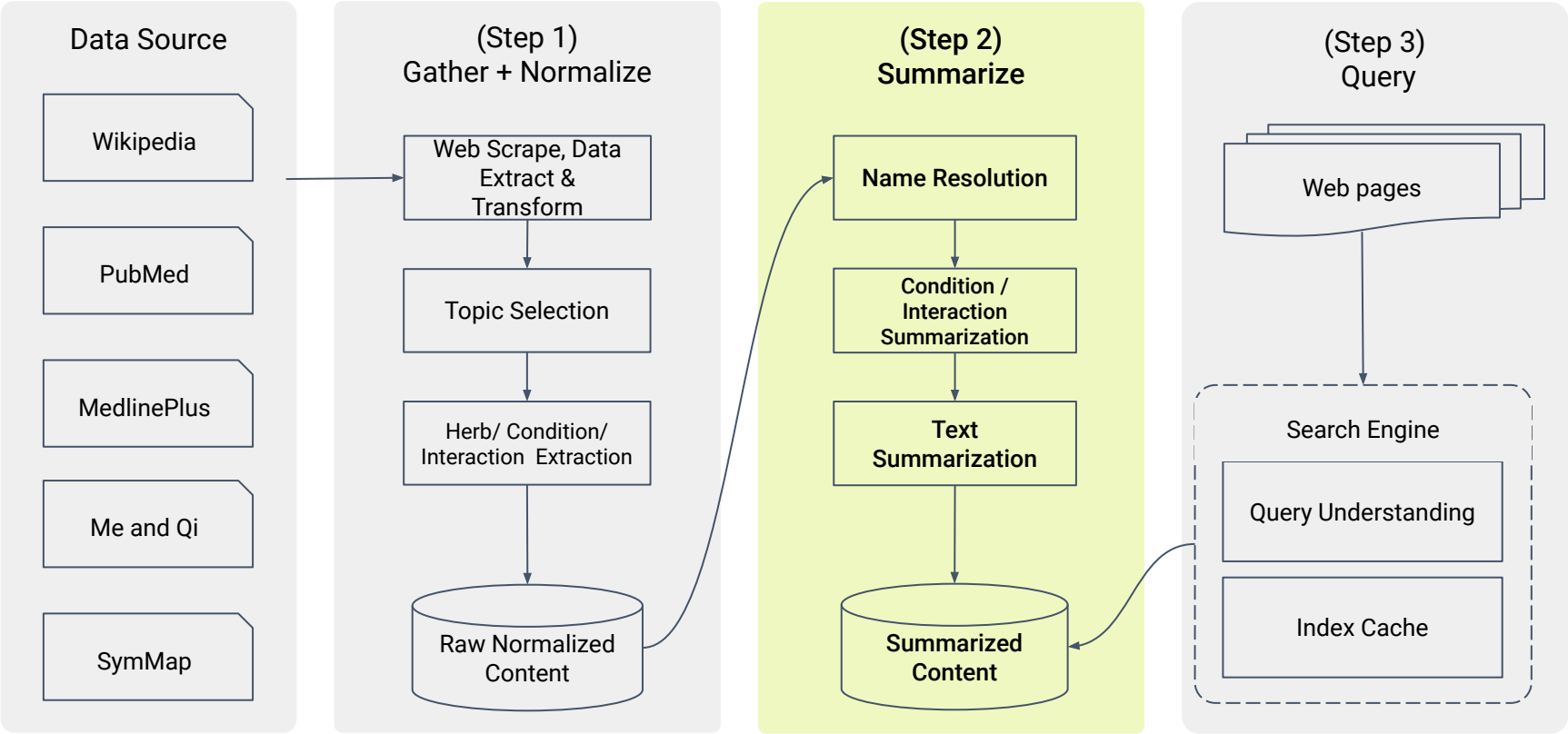


Learnings:

General, configurable pipeline that can be measured at each stage and used across multiple problems and domains.



(STEP 2) SUMMARIZE



HERB NAME RESOLUTION

Same herb can have many different names

Data Source	Name
Me and Qi	Dittany root bark
SymMap	Dictamni Cortex
Wikibooks	Densefruit Pittany Root-Bark
Other names	White Fresh Bark
Other names	Burning Bush Root
Other names	Dictamnus dasycarpus root-bark

Chinese: 白鲜皮

Pinyin: Bái Xiān Pí

CONDITION NORMALIZATION PROCESS

Source	Condition
Me and Qi	Dysmenorrhea
SymMap	Painful Periods
Wikipedia	Menstrual Cramps Dysmenorrhea

USER FEEDBACK

USER FEEDBACK

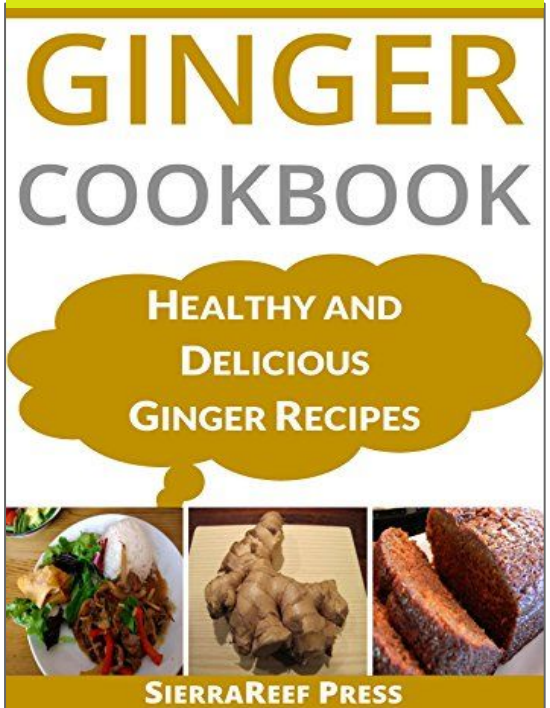
George Shea (Kaiser Santa Clara Pharmacy Director for 13 pharmacies + TCM user): “This is **immediately practical for the average layperson** and overcomes the typical resources that are **overly technical**. Having cross references in itself is a major step forward in the field...I’ve seen **many of these herbs in the TCM shops** and am confident you hit the 80 in the 80-20 distribution of herbs.”

USER FEEDBACK

Feedback	Example/Description	Team Response
Some of the medical names are too technical	E.g. hyperlipidemia vs. cholesterol	Manually mapped back some of technical names to “lay” names
Where can I find all the herbs/conditions?	Can only lookup herb through search or through hyperlinks, need prior knowledge	Added landing pages for master list of each
I can go from an herb to condition but not back	Hyperlinks on herb page but not conditions page	Added hyperlinks to conditions page
Looks like there are some typos in the summaries	E.g. Hanging empty parentheses, or new lines between sentences	Used regex to clean up some of the common “junk” patterns

STILL MORE TO DO

COOKING RECIPES



RESEARCH EVALUATION

DRUG INTERACTIONS AND NORMALIZATION

RxNav

Navigating RxNorm Drugs

WRAP UP

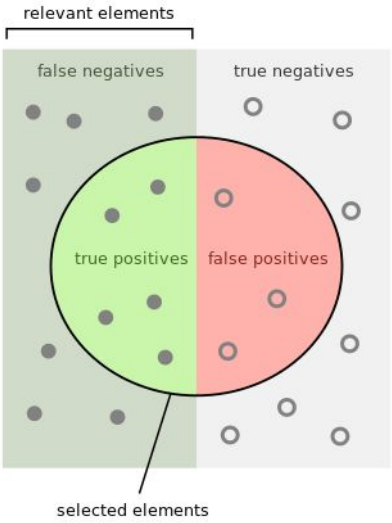
HERBERT

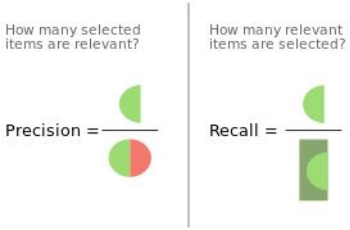
QUESTION & ANSWER

APPENDIX

RELEVANT TOPIC SELECTION EVALUATION

- Metrics: Precision, Recall, F1
  - Want to see how many relevant topics selected
  - Due to **high imbalance in content topics**, preferred over accuracy (**average 2/6 are potentially condition-related**)
  - Data: [‘50 fundamental herbs’]
  - Agreement b/w annotators: **Fleiss Kappa of .65** (“substantial” but not perfect)





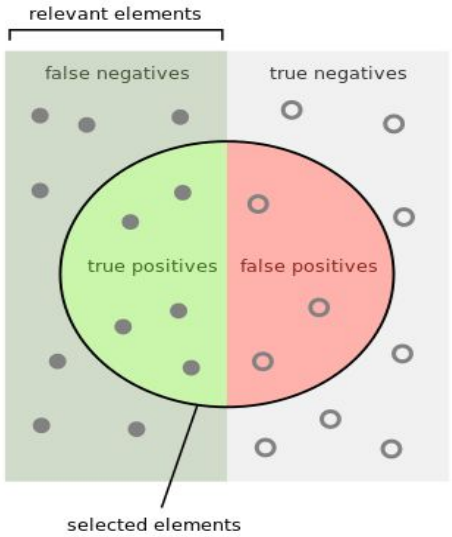
Precision	Recall	F1
92%	63%	75%

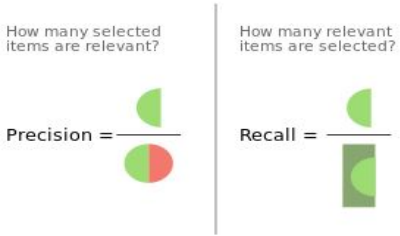
CONDITION EXTRACTION EVALUATION

- **Metrics:** Precision, Recall, F1
- 75 entries across 5 Herbs
- Aimed to maximize precision for reliability

Precision*	Recall*	F1*
94%|92% PA	68% | 77% PA	79%| 84% PA

*Deflated/Pessimistic since we remove entries coming only from a single source that's less trusted (e.g. Wikipedia only).
PA = "Post-adjustment"





CONDITION EXTRACTION EVALUATION BY SOURCE

Source	Precision	Recall	F1
NIH	91%	48%	63%
Pubmed	100%	75%	86%
Wikipedia*	86%|80% PA	40%|86% PA	55%|83% PA
Me & Qi	97%	94%	96%

*Deflated/Pessimistic since we remove entries coming only from a single source that's less trusted (e.g. Wikipedia only).
PA = "Post-adjustment"

Appendix :NAME ENTITY EXTRACTION EVALUATION

Software Package	Processing Times Per Abstract (ms)	Sentence (ms)
NLP4J (java)	19	2
Genia Tagger (c++)	73	3
Biaffine (TF)	272	29
Biaffine (TF + 12 CPUs)	72	7
jPTDP (Dynet)	905	97
Dexter v2.1.0	208	84
MetaMapLite v3.6.2	293	89
en_core_sci_sm	32	4
en_core_sci_md	33	4

Dataset	Baseline	SOTA	+ Resources	**sci_sm**	**sci_md**
BC5CDR (Li et al., 2016)	83.87	86.92b	89.69[bb]	78.83	83.92
CRAFT (Bada et al., 2011)	79.55	-	-	72.31	76.17
JNLPBA (Collier and Kim, 2004)	68.95	73.48[b]	75.50[bb]	71.78	73.21
BioNLP13CG (Pyysalo et al., 2015)	76.74	-	-	72.98	77.60
AnatEM (Pyysalo and Ananiadou, 2014)	88.55	91.61[**]	-	80.13	84.14
BC2GM (Smith et al., 2008)	84.41	80.51[b]	81.69[bb]	75.77	78.30
BC4CHEMD (Krallinger et al., 2015)	82.32	88.75[a]	89.37[aa]	82.24	84.55
Linnaeus (Gerner et al., 2009)	79.33	95.68[**]	-	79.20	81.74
NCBI-Disease (Dogan et al., 2014)	77.82	85.80[b]	87.34[bb]	79.50	81.65

bb: LM model from Sachan et al. (2017) **b**: LSTM model from Sachan et al. (2017)
a: Single Task model from Wang et al. (2018) **aa**: Multi-task model from Wang et al. (2018)
** Evaluations use dictionaries developed without a clear train/test split.

Table 5: Test F1 Measure on NER for the small and medium scispaCy models compared to a variety of strong baselines and state of the art models. The **Baseline** and **SOTA** (State of the Art) columns include only single models which do not use additional resources, such as language models, or additional sources of supervision, such as multi-task learning. **+ Resources** allows any type of supervision or pretraining. All scispaCy results are the mean of 5 random seeds.

- Compared to state of the art ("SOTA"), 2x the speed , 1/10th size, at the cost of about 7% difference in F1

TEXT SUMMARIZATION EVALUATION PLAN

- Metrics: Rouge-N Family
  - Like recall
  - Overlap of summary and “golden summary”

- Data: Plan to annotate and summarize a “train”, “test”, and “validation set”
  - Measure agreement b/w annotators

- ROUGE-N

$$ROUGE-N = \frac{\sum_{S \in S_H} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in S_H} \sum_{g_n \in S} C(g_n)}$$

where

S_H is the set of manual summaries

S is an individual manual summary

g_n is an N-gram

$C(g_n)$ is the number of co-occurrences of g_n in the manual summary and automatic summary

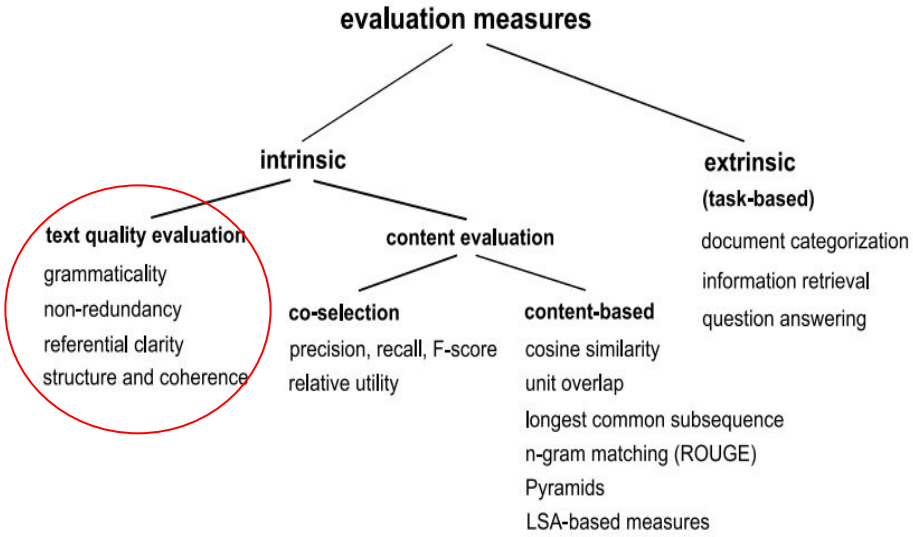
TEXT SUMMARIZATION EVALUATION

Need annotators to rate on qualitative scale since we have length-constrained extractive summary. Task strongly time intensive for humans.

Get measurement of agreement.

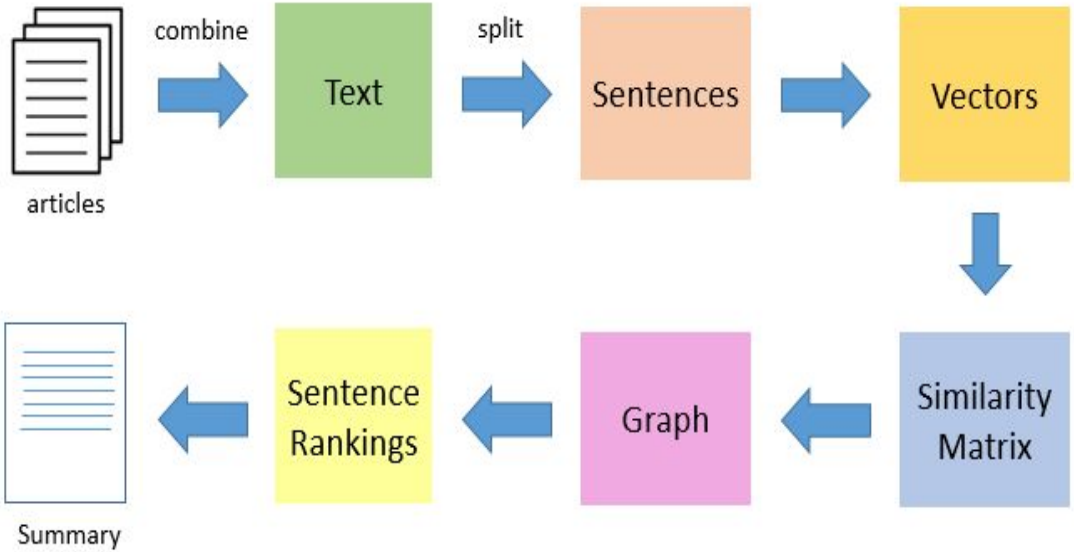
Not done yet due to time constraint.

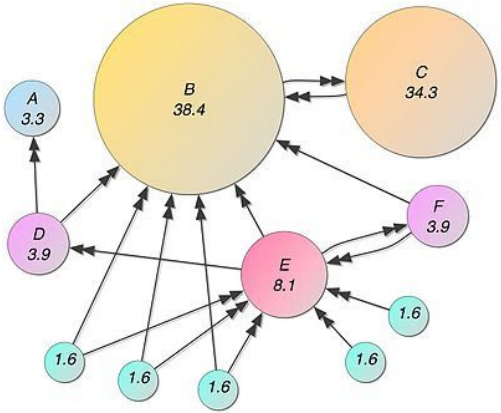
For now at least from our own QA and SMEs , no major complaints...



Reference: “EVALUATION MEASURES FOR TEXT SUMMARIZATION” by Josef Steinberger & Karel Je˘zek

Appendix: TEXT SUMMARIZATION - TEXT RANK





Key Concept: Use Pagerank to score sentences with the assumption that “summary sentences” are similar and central to most other sentences.

Appendix: Website Index (Lookup Tags) V1

- Don't want to comb every document every time we query so we build indices
- Key indices: Herbs, Conditions
- For each key index produce posting: (stemmed_word, [(documents containing word, length)], frequency stats)
- E.g. 'ginger' appear in 5 documents total

  Posting:
  ('ging', [('Ginger', 150), ('Ginger Root', 165), ...
  ('Upset Stomach', 80)...], 100)

Appendix: Matching/Relevance Function V1

Okapi Best Matching-25 (BM-25)

Document (indices)

Reward rare words/Discount common words

Reward word matches

Discount short documents that don't match at all

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$

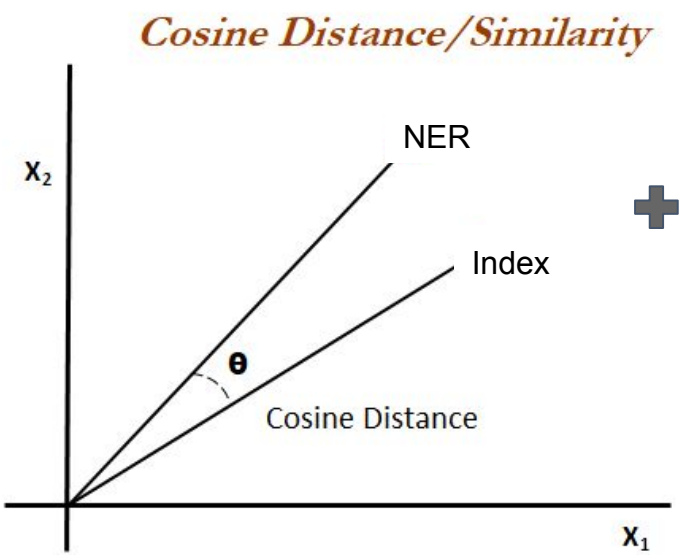
Query words (q_1 … q_n)

Discount long documents/Reward short documents

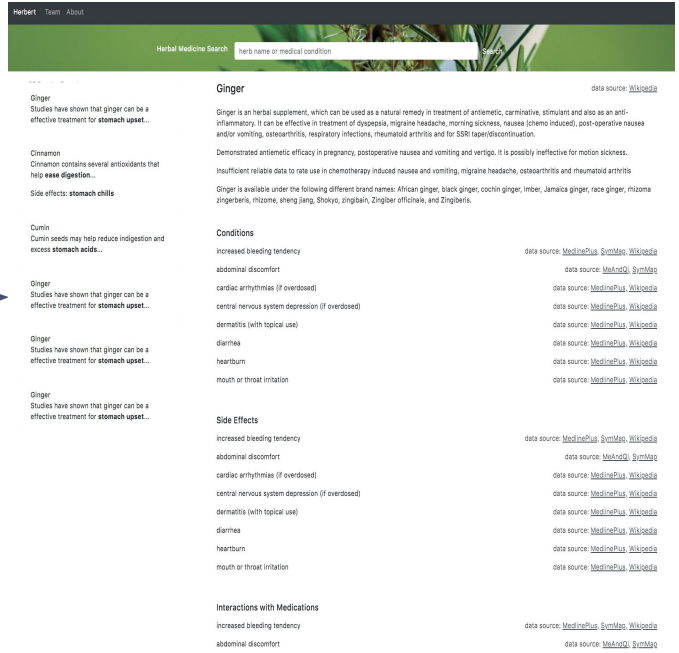
Website: Index and Matching V2

Query: “Does ginger help with stomach pain?”

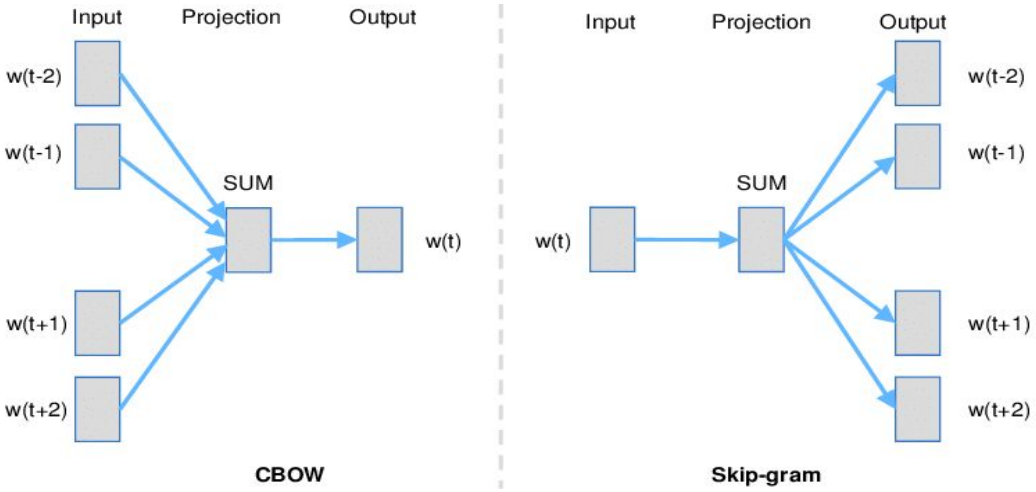
NER: ginger, help, stomach pain

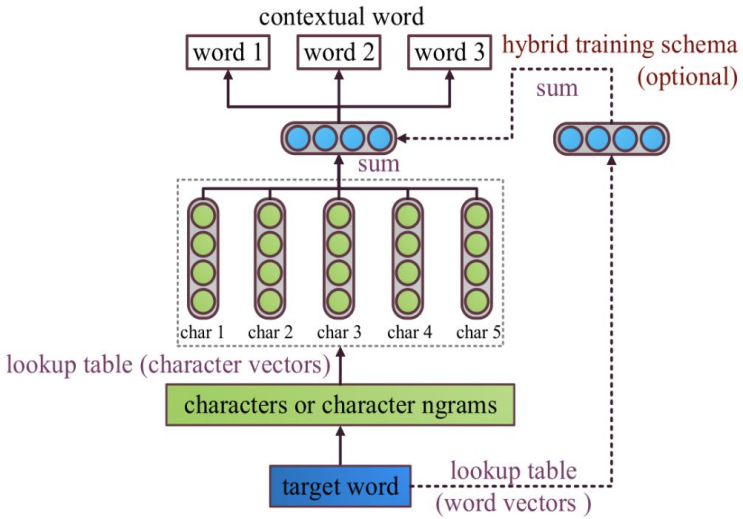


+ **Thresholding** →



Appendix: RELEVANT TOPIC SELECTION



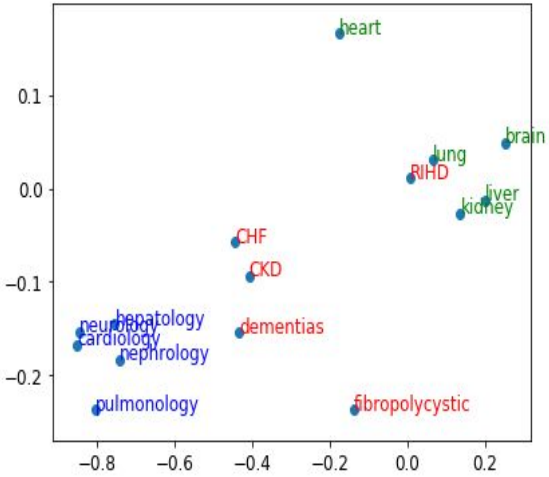


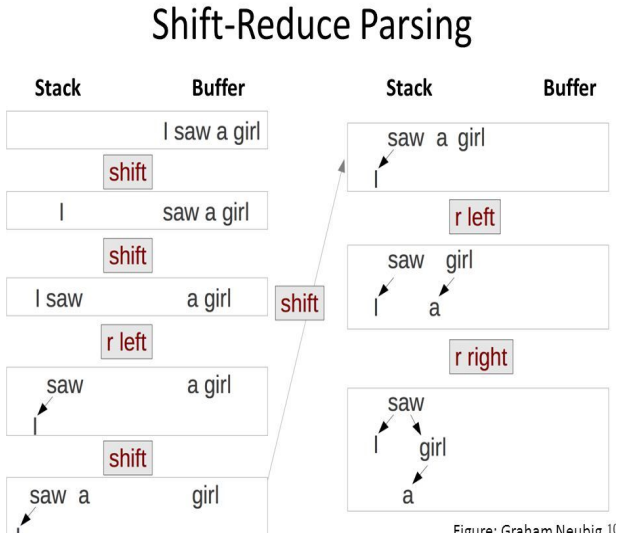
- **Word vectors capture some meaning** via distributional hypothesis
- Incorporating **character n-gram** vectors can help **capture structure**.
- **EX)** n=4 for word “toxic”:

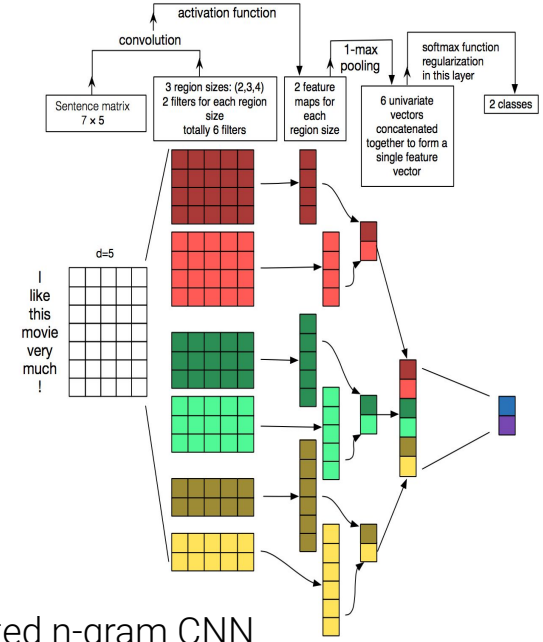
  <tox/toxi/oxic/xic>

  Where “<” denotes start, “>” denotes end
- Now **have way of making sense of out of vocab words** like “toxicology”!

Appendix: NAME ENTITY EXTRACTION

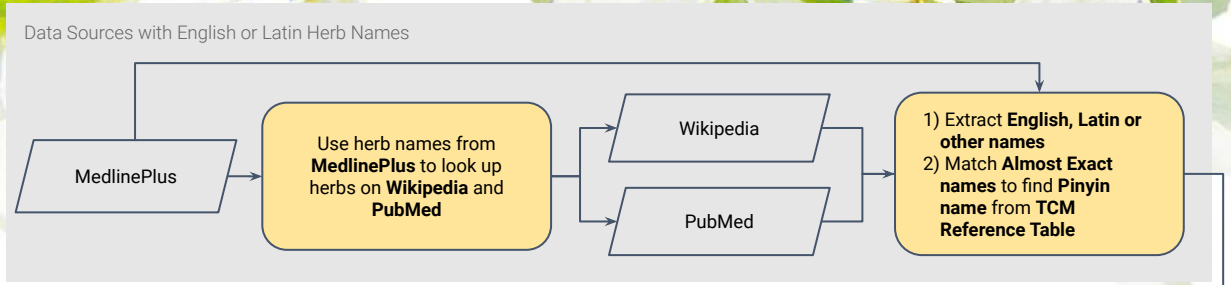






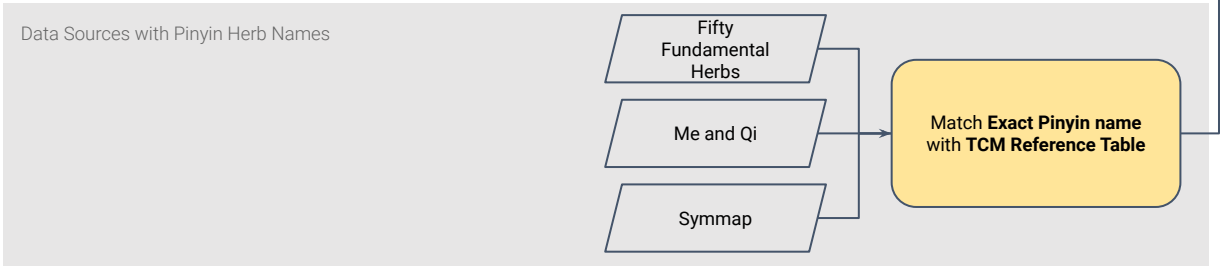
- **Guts of model:** Hashed Word Embeddings + Parsing Features + Dilated n-gram CNN
- **Rough summary**: Look at meaning of words + grammatical grouping structure to determine if entity
- **Why not LSTM/CRF**: Speed of prediction, relative training size needed, relative simplicity of features

Appendix:Herb Name Resolution Process



% of herbs found in ref. table
MedlinePlus: 61%
Wikipedia: 61%

Almost Exact means when the reference table herb name contains all the words (without stop words and no specific order) of the herb name



% of herbs found in ref. table
Fifty Fundamental Herbs: 74%
Me and Qi: 94%
Symmap: 86%

TCM Reference Table:
Harvard TCM-ID Table has ~10000 TCM herb names and IDs

Match **Exact Pinyin, or Almost Exact English, Latin, or other names** among data sources

% of herbs found in ref. table or among other data sources

MedlinePlus: 65%*
Wikipedia: 65%*
Fifty Fundamental Herbs: 86%
Me and Qi: 97%
Symmap: 89%

***Note**: MedlinePlus and Wikipedia are ~98% matched because we use MedlinePlus to guide Wikipedia herb search

Use **BioBERT word embeddings** to generate Pinyin name and find match from **TCM Reference Table**

Work-in-progress

Other potential Sticky facts to use (placeholder)

Health is the third most searched thing on internet
1. Email
2. Online Shopping
3. Healing

Google receives more than 1 billion health questions every day

Jackie Drees - Monday, March 11th, 2019 Print | Email

SHARE Tweet Share 4

An estimated 7 percent of Google's daily searches are health-related, according to Google Health Vice President David Feinberg, MD, *The Telegraph* reports.

If We Could Do It Again...

- True to namesake: Use Bert Embeddings (Sequence to sequence rather than just words)
- Medical domain relation extraction system

IMPACT

Bring clarity and reduce search fatigue.