Articles::

<https://arxiv.org/pdf/1801.10198.pdf> (using Wikipedia articles as summaries, references as originals)

<https://arxiv.org/pdf/1301.3781.pdf> (king - man + woman = queen)

<https://arxiv.org/pdf/1706.03762.pdf> (use of attention networks for generating sentences)

<https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/> (tutorial for use of RNN to generate text)

<https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/> (extractive vs abstractive seems to be a key distinction)

<https://code.google.com/archive/p/word2vec/>

<https://machinelearningmastery.com/encoder-decoder-deep-learning-models-text-summarization/>

<https://machinelearningmastery.com/encoder-decoder-models-text-summarization-keras/>

More abstract and mathematical:

<https://www.stat.fi/isi99/proceedings/arkisto/varasto/amar0019.pdf>

<https://arxiv.org/pdf/1303.0818v5.pdf>

<https://arxiv.org/pdf/1306.0514v4.pdf>

<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>

Questions:

What is our project?

Text summarisation, creating more concise versions of an input document that retain the core concepts. Allows improved efficiency

Further research may be needed

Reddit smmry TL;DR tracks noun frequency ranking, locates sentences, outputs result.

TF-IDF to identify keywords/phrases

Word2vec vector encoding to generate paragraph vectors, produce more concise versions of the same vector, Doc2vec.

For sentence generation, consider RNNs, attention networks

What are the merits of each?

What resources do we need?

Training dataset - want large number of examples of documents with accompanying summaries

Use wikipedia? Restrict to subset of articles? Try to just replicate headers

<https://en.wikipedia.org/wiki/Wikipedia:Database_download>

Use ArXiv papers, abstracts as summaries

<https://arxiv.org/help/bulk_data>

Search TL:DR for more broad class of summaries, informal register

<https://www.reddit.com/r/datasets/comments/65o7py/updated_reddit_comment_dataset_as_torrents/>

eBooks with synopses - Project Gutenberg for the books, need to find synopses somewhere

<https://github.com/puntonim/gutenberg-bulk-downloader>

Which milestones do we need to reach?

Get data extraction working

Get extractive method working

Get paragraph -> vector -> paragraph working

Start working on individual paragraphs -> sentences

Put together multiple paragraphs -> paragraphs

Study RNNs and attention networks