Annotated Bibliography:

Pitch 2: Using Machine learning algorithms in order to generate a query relevant summary, by taking the ideas in the text, categorizing them and highlighting the most important sentences so that the user can refer back to the original text and without reading the whole thing extracting the most important pieces of information.

Text summarization using Latent Semantic Analysis: Journal of Information Science

```
@article{ozsoy2011text,
  title={Text summarization using latent semantic analysis},
  author={Ozsoy, Makbule Gulcin and Alpaslan, Ferda Nur and Cicekli, Ilyas},
  journal={Journal of Information Science},
  volume={37},
  number={4},
  pages={405--417},
  year={2011},
  publisher={Sage Publications Sage UK: London, England}
}
```

- There have been several different approaches for text summarization throughout history. The first one is based in frequency of the words in a document. There have been many studies based on simple features for summary analysis using terms from keywords/key phrases, terms from user queries, frequency of words, and position of words/sentences.
- The paper introduces two more projects to look at when dealing with summarization specifically using statistical analysis. Is a concept where relevant information is extracted from dictionaries and WordNet and used together with natural language-processing methods. Other useful methods to look at are Text connectivity and graph-based summarization approaches.
- The most important aspect of this paper is the information provided in regards with Machine Learning, the approach I have in mind in order to proceed with the pitch. The text pointed me to my next resource that uses algorithms such as Naı¨ve-Bayes, Decision Trees, Hidden Markov Model, Log–linear Models, and Neural Networks.
- The text proposes an approach that I could use to measure the effectives of my summary by comparing the results I get with the Latent Semantic Analysis. They use an algebraic-statistical method that extracts hidden semantic structures of words and sentences. It is an unsupervised approach that does not need any training or external knowledge. The strategy is called LSA, it uses the context of the input document and extracts information such as, which words are used together and which common words are seen in different sentences.
- This paper also provided important information of the way of measuring the accuracy of each summary by using the ROUGE evaluation approach which is based on n-gram co-occurrence, the longest common subsequence and the weighted longest common subsequence between the ideal summary and the extracted summary. The n-gram based ROUGE score, ROUGE-N, is based on comparing n-grams in the ideal summaries and the reference summary.

A Survey on Automatic Text Summarization:

```
@article{tas2007survey,
  title={A survey automatic text summarization},
  author={Tas, Oguzhan and Kiyani, Farzad},
  journal={PressAcademia Procedia},
  volume={5},
  number={1},
  pages={205--213},
  year={2007}
}
```

- The paper presented different approaches using known Machine Learning algorithms, including their advantages and the way to categorize the sentences in a text, for example: Naive-Bayes Methods where each sentence was given a score, and only the n top sentences were extracted.
- A second important note from this text is the strategy introduced while using decision trees. The "position method", arises from the idea that texts generally follow a predictable discourse structure, and that the sentences of greater topic centrality tend to occur in certain specifiable locations.
- New ideas as of how to test were also valuable from this paper, starting by using previously unseen text was used for testing whether the same procedure would work in a different domain. The first evaluation showed contours exactly like the training documents. In the second evaluation, word overlap of manual abstracts with the extracted sentences was measured.
- Two important points from this text, the neural net investigation has much more to explore now that tools have improved and increased their accuracy. The paper briefly mentions the possibility of using Neural Nets in past papers, but the research method and actual experiments are relatively poor in contrast with the other ones. Second important thing I could take from this text is the fact that there are different options while summarizing, including Multi-Document Summarization, and Single-Document Summarization.
- In terms of evaluation this paper raises the same method as in the other one which makes me think is one of the most valid approaches when evaluating the accuracy of a summary. Lin (2004) introduced a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE)23 that have become standards of automatic evaluation of summaries.

Automatic Text Summarization Using a Machine Learning:
Approach Joel Larocca Neto, Alex A. Freitas, and Celso A. A. Kaestner

- The paper has important insights in how to break down a text summary: Their research is divided in the following three steps:
    - (1) in the preprocessing step a structured representation of the original text is obtained;
    - (2) in the processing step an algorithm must transform the text structure into a summary structure;

    - and (3) in the generation step the final summary is obtained from the summary structure.
- This paper also clarified the specific approach I have in mind of the highlighting software. As I will be using the exact same phrasing without changing the meaning the approach my software will take would be a **deeper approach**, which assume a semantics level of representation of the original text.
- A second way of testing without using the ROUGE method is introduced in this paper when testing Keynes Classifier. They did two series of experiments: in the first one, they employed automatically produced extractive summaries; in the second one, manually produced summaries were employed. In all the experiments they have used a document collection available in the TIPSTER document base.
- Text summarization is important as it avoids the problem of subjective evaluation of the quality of a summary, which is a central issue in the text summarization research.

Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis: Yihong Gong, Xin Liu

- There are two types of summary generation: a generic summary and a query-relevant summary. A generic summary provides an overall sense of the document's contents. A good generic summary should contain the main topics of the document while keeping redundancy to a minimum. On the other hand a query-relevant summary is essentially a process of retrieving the query relevant sentences/passages from the document.
- I will dig deeper into the method proposed by the University of southern California: The SUMMARIST text summarizer from the University of Southern California strives to create text summaries based on the equation: summarization=topic identification+ interpretation +generation.
- The article proposes two methods. Both, first decompose the document into individual sentences, and creates a weighted term-frequency vector for each of the sentences.
- **Relevance Measure:** Decompose text into individual sentences, compute the relevance score of each sentence with the whole document. Then select the sentence k that has the highest relevance score, and add it to the summary. Once the sentence k has been added to the summary, it is eliminated from the candidate sentence set, and all the terms contained in k are eliminated from the original document.
- **Latent Semantic analysis:**
    - First I will need to understand what is singular value decomposition (SVD).
    - The process starts with the creation of a terms by sentences matrix with each column vector representing the weighted term-frequency vector of sentence i in the document under consideration.
    - Words with similar motifs or used in similar contexts, will be mapped near to each other in the r-dimensional singular vector space.
- The testing technique was based on a database of news articles consisted of closed captions of 549 news stories whose lengths are in the range of 3 to 105 sentences.

Dynamic Coreference-Based Summarization. Breck Baldwin, Thomas S. Morton

- The summarization technique was developed with the CAMP NLP framework.
- Associations between tokens in the query, headline, and the body of the document. Event references arc captured by associating verbs or nominalizations in the query with verbs and nominalizations in the cloc1nnent
- The scores of each sentence from 1-6 is ranked in descending order while the 7 is ranked in ascending order. The top-ranked sentence is selected 1 and scores 1, 3, and 5 are recomputed in order to select the next sentence. Selection halts when all coreference chains in the query have be covered and the summary contains at least 4 sentences.

Natural language processing for information retrieval
David D. Lewis Karen Sparck Jones

- This papers goes in depth about the NLP strategies to categorize data. Retrieval depends on indexing, that is on some means of indicating what documents are about. Indexing requires an indexing language with a term vocabulary and a method for constructing requests and document descriptions.
- In the big scale of things categorizing data could be both controlled-language indexing and more sophisticated natural-language indexing imply nontrivial NLP, so the other issue is whether the required NLP capabilities are available or in prospect, since large-scale human full-text processing is not a practical proposition.
- The paper bases the explanation into three:
    - The “words,” “phrases,” and “sentences” that form individual document descriptions and express the combinatory, syntagmatic relations between single terms captured by the system’s NLP-based text-processing apparatus;
    - The “classificatory” structure over the document file as a whole that indicates the paradigmatic relations between terms and allows controlled term substitution in NLP-based indexing and searching; and
    - The system’s NLP-based mechanisms for searching and matching.
- Word have to undergo a process of normalization in order to be processed the paper proposes two useful approaches to normalize words either by Semantic normalizations and statistical associations.
-

Pitch 1: Using Neural Nets to train a Machine Learning model to generate an accurate guess for the genre of different songs.

MUSICAL GENRE CLASSIFICATION USING SUPPORT VECTOR MACHINES
Changsheng Xu, Namunu C. Maddage, Xi Shao, Fang Cao, Qi Tian

- The very first thing in the exploration of music genre identification, is having a way of measurement for the computer to identify what is going on. Beat spectrum is a measure to automatically characterize the rhythm and tempo of the music. It is achieved by three different steps: First, the music is parameterized using a spectrum or other representation. This results in a sequence of feature vectors. Second, a distance measure is used to calculate the similarity between all pairwise combinations of feature vectors.

The obtained similarity is embedded into a two-dimensional representation called similarity matrix.

- The genre classification itself was done through a multi-layer classifier based on SVM is used to discriminate musical genres. SVM is a useful statistic machine learning technique that has been successfully applied in the pattern recognition area. The basic idea is to transform input vectors into a high dimensional feature space using non-linear transformation, and then to do a linear separation in feature space.
- This paper gage me a scope of how much data could be required to make a valuable analysis of the music genres and to scope down the software by making it identify small sets of genres. In this study they used 4 different types of genres and 60 music samples as training data including 15 pop music, 15 classic music, 15 rock music and 15 jazz music. Each sample is segmented into 2000 frames and the length of each frame is 882 sample points.

Neural Network Music Genre Classification Classification des genres de musique par réseau neuronal

- In the research of finding what the best approach is when using Neural nets this paper suggested important insights on the type of neural net I would have to use: A convolutional Neural Net (CNN). CNN is a type of neutral network that is intended to process multidimensional vectors such as images
- This paper is amazing, it provides insights of the way that Shazam classifies music and also their attempts in classifying genre. Shazam uses a technique they refer to as a song's signature. Shazam defines a song's signature as the large peaks in amplitude taken from the song's spectrogram
- The NN had two layers in order to identify genre
- Another approach that complements the NN is the Spectrogram. He used spectrogram snippets as input into his CNN. The architecture of the CNN consisted of four CNN layers, a fully connected layer, and a softmax function to classify the results into the genre classes specified. The test set accuracy was not provided, however, the reported validation set accuracy was 90%. Validation accuracy is a measure of the NN's accuracy after each epoch using data the NN has not seen before
- An important take away from this paper is that their results changed because they trained their Neural Net with different amount of music, but their results once they used the same amount of songs per genre As mentioned above, the number of songs in each genre was unequal in the first data set and this was maintained in the second data set for consistency and for comparing test accuracy. The second data set had enough songs in the library to form an equal representation of songs across all genres.
- I could use the suggested section for improved work to choose what part of my research I will be focusing. Future research work with respect to modifications to the algorithms and feature engineering includes changes to the initialization of the weights and experimenting with different filters while converting the mp3 file to a spectrogram.

Music Genre Classification
Michael Haggblade Yang Hong Kenny Kao

- This papers showed a reliable database of songs where they extracted all the information they used. Marsyas (Music Analysis, Retrieval, and Synthesis for Audio Signals) is an open source software framework for audio processing with specific emphasis on Music Information Retrieval Applications.
- In order to communicate with the computer the approach could be MFCCs, a way to represent time domain waveforms as just a few frequency domain coefficients

Steps used in this paper:

1. multiply by a hamming window to smooth the edges
2. Fourier Transform to get the frequency components
3. map the frequencies to the mel scale, which models human perception of changes in pitch, which is approximately linear below 1kHz and logarithmic above 1kHz
4. calculating triangle window coefficients based on the mel scale, multiplying that by the frequencies, and taking the log
5. Discrete Cosine Transform, which serves as an approximation of the Karhunen-Loeve Transform
6. first 15 of these 20 frequencies since higher frequencies are the details that make less of a difference to human perception and contain less information about the song

- Again he K-Neiarest Classifier method comes up might be a good idea to use it as a possible new approach when classifying genres