NLP:

* Technical:
  + Sentiment analysis [logistic regression or naïve bayes [for spam filtering too] for binary classification]: extract useful features [process in training sets [remove punctuation stemming words and tokenizing], create frequency dictionary, use stemming and stop words], train classifier [cost-function: y is training example and h is prediction] and make predictions

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

* + Vector space models: used in identifying similarity [ie in summarisation]. In information extraction [who, why and where]. Similarity using cosine similarity and Euclidean distance
  + PCA: extract uncorrelated features [dimensionality reduction]. Eigenvectors and eigenvalues from covariance matrix [perform SVD, eigenvalues and eigenvectors]
  + K-nearest neighbours and hash values and functions [locality sensitive hashing to reduce computational cost of k-nearest neighbours]
  + Autocorrection and web search suggestions: Calculate minimum edit distance [insert, delete, and replace for edit cost]. Dynamic programming 🡪 Levenshtein distance + backtrace. Markov chains [initial and transition probability distributions]. Viterbi algorithm [graph algorithm] for speech tagging and speech recognition
  + N-gram probability model and sequence models [Katz backoff for smoothing]

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated

* + Word embeddings: one-hot vectors (0, 1), word embedding vectors (in real world embeddings in hundreds of dimensions) word2vec Global Vectors (GloVe) fastText BERT ELMo GPT-2. CBOW model
* Applications:
  + Business
    - Sentiment analysis for brand and market research [analyse customer reviews, comments and social media mentions / convert and classify words and phrases into viable marketing insights]
    - NLP chatbots [customer service automation]. Ability for personalised conversations to bring in human employees [look at BERT and XLNet]
    - Build an overview of a company’s unique competitive landscape [demographics and size-up competitors]
    - Automate report generation [speech-to-text dictation and formulated data entry]. Target relevant information in unstructured text
  + Healthcare
    - Automate administrative workflows [replace type writing with voice notes]
    - Unstructured clinical record and patient feedback [pull key word and phrases from free text]
    - Value based care model in healthcare: data rich health system to analyse post-care survey feedback and other sources of unstructured text
    - Predictive analytics [more complicated and used to identify subsets of geographic regions ethnic groups etc with different types of health disparities]