Summary of “Bag of Words” text processing:

--What’s my end goal?  The end goal of text analysis, for our purposes, is to analyze a domain-specific piece of unstructured text (it can be one article, or even a collection of articles on a specific topic) to find a list of the most important concepts in the unstructured text., along with each concept’s frequency of appearance in the unstructured text  We determine if a concept is important to the subject based on frequency – how often does that concept appear in the unstructured text?  This can be a very powerful tool that you are building, which can help you analyze a topic!

--What’s my approach?  You want to take the unstructured text block, scrub the data to make the text standardized, extract the concepts, and count each concept’s frequency to determine concept importance to the topic at hand.

--What is the process?

\*To scrub the data, you perform:

   \*\*Joining compound concepts == compound concepts are phrases comprised of multiple words.  When the individual words are put together, the individual words’ meaning changes.  “First aid”, or “civil war”, are two examples of compound concepts.  Join the words that comprise the compound concepts together with an underscore; eliminate the whitespace.  For example, replace “first aid” with “first\_aid” to preserve the components that make up the compound concept.

   \*\*Perform normalization == standardize your text.  Replace similar concepts with a single synonym – for instance, replace “maroon”, “dark red”, and “deep red” with “maroon”.  This will give you a more accurate frequency count.  Spell out acronyms – for instance, replace “USA” and “America” with “United\_States\_of\_America”.  Deal with contractions.

   \*\*Perform stemming == get to the root of your words.  Make nouns singular and replace verbs with their present tense forms.  This also improves your frequency count.

   \*\*Noise / stop words removal == replace any non-meaningful words with whitespace once you preserve your compound concepts with underscore insertion.  This can include pronouns, prepositions, verbs of being…

\*Once you scrub/standardize your text, pull out the concepts and count each concept’s frequency.  Frequency is a proxy for importance.  Sort your concept list by frequency so the most important concepts float to the top of the list.

--What does this give me?  By creating this programming tool, you can then analyze any domain-specific topic text to determine the core ideas in the text.  Note that your tool is not article-specific; it is domain specific.  You can analyze any unstructured text with your tool!  It’s very powerful!  You can also process multiple pieces of unstructured text at once, finding the common, important concepts across multiple items.