Dependencies: We have several dependencies on outside libraries:

- Stanford NLP kit: in particular, we use the simple package of the kit
- Jsoup: we use Jsoup to parse HTML into plaintext
- Porter's Algorithm: This is Porter's implementation of his own algorithm. We added the static `stemString` method as an easy interface for our use.

Preprocessing: As before, we are using Jsoup to parse the original HTML. We further process this via regex find/replace to remove citation numbers like this [1].

Parsing: We used the Stanford NLP library for our parsing. We create a new Document (Stanford object) from the entire preprocessed text, and let the parser decide where the sentence splits are. We process each Sentence (Stanford object) into a Statement (our object). A Statement has a subject, a verb, and a list of objects. They are identified from the Sentence as follows:

- The verb is the token which is the root of the sentence. We found this assumption to almost always be true.
- The subject is the token which has an incoming dependency relationship of "nsubj" and is governed by the root token.
- An object is any other token which is governed by the root, filtering out punctuation and conjunctions.

All words are stored ignoring case, and stemmed using Porter's algorithm. Note that there are some cases which the stemmer does not handle in an ideal way, but at least it will handle them in the same irregular way on both the document side and the user query side.

Here are some example successful conversions of Sentences to Statements:

```
Original: Lincoln only vetoed four bills passed by Congress; the only
important one was the Wade-Davis Bill with its harsh program of
Reconstruction.
Subject: lincoln
Verb: veto
Objects: onli, bill

Original: The law assigned land for a lease of three years with the
ability to purchase title for the freedmen.
Subject: law
Verb: assign
Objects: land, leas
```

Original: Confederate Vice President Stephens led a group to meet with Lincoln, Seward, and others at Hampton Roads.
Subject: stephen
Verb: led
Objects: meet, group

Original: Grant waged his bloody Overland Campaign in 1864.
Subject: grant
Verb: wage
Objects: 1864, campaign

Original: His children, including six-year-old Thomas, the future president's father, witnessed the attack.
Subject: children
Verb: wit
Objects: attack

Here are some example of unsuccessful translations:

Original: The case is famous for Lincoln's use of a fact established by judicial notice in order to challenge the credibility of an eyewitness.
Subject: case
Verb: famou
Objects: is, us

Original: He was the first president from the Republican Party.
Subject: he
Verb: presid
Objects: the, wa, first, parti

Occasionally the root is not a verb, especially when the verb is a past-tense conjugation of "to be". The parser appears to treat it as if it is a passive voice statement, as in "He was told to do his work." The silver lining is that user input will be mishandled in the same way.

Original: Having composed the Proclamation some time earlier, Lincoln had waited for a military victory to publish it to avoid it being perceived as the product of desperation.
Subject: lincoln

```
Verb: wait
Objects: victori, had, compos
```

Given this subject and verb, we probably only really wanted "victory" to be the object.

Querying: The user inputs a phrase, which is parsed into a Sentence, then a Statement. We check if the user's Statement matches any in the document. To match, they must have the same subject and verb, and a nonempty intersection of objects. If a match is found, the original text of the document's statement is returned to give the user context.

Example successful queries:

```
> the family moved west
Statement verified. Original source: In early March 1830, fearing a
milk sickness outbreak along the Ohio River, the Lincoln family moved
west to Illinois, a non-slaveholding state.

> the lincoln family moves north
Statement verified. Original source: In 1816 the family moved north
across the Ohio River to Indiana, a free, non-slaveholding territory,
where they settled in an "unbroken forest" in Hurricane Township,
Perry County.

> the family moved south
Statement could not be verified.

> lincoln died in april
Statement verified. Original source: After remaining in a coma for
nine hours, Lincoln died at 7:22 am on April 15.

> lincoln died in may
Statement could not be verified.

> lincoln became a lawyer
Statement verified. Original source: Lincoln became an able and
successful lawyer with a reputation as a formidable adversary during
cross-examinations and closing arguments.
```

Challenges/shortcomings:

Preprocessing does not completely clean the text. For example, the phonetic pronunciation at the beginning of the article doesn't parse quite right, and we get `k?n/ (February 12, 1809 &#65533; April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865.` as a sentence:

```
> k?n/ was president
Statement verified. Original source: k?n/ (February 12, 1809 &#65533; April 15, 1865) was the 16th President of the United States, serving from March 1861 until his assassination in April 1865.
```

There is no translation of pronouns to the nouns to which they refer. We could not find a practical solution to this, especially as sometimes the referenced noun was not in the same sentence as the pronoun:

```
> lincoln believed in god
Statement could not be verified.

> he believed in god
Statement verified. Original source: However he did believe in an all-powerful God that shaped events and, by 1865, was expressing those beliefs in major speeches.

> lincoln was president
Statement could not be verified.

> he was president
Statement verified. Original source: He was the first president from the Republican Party.
```

There is no synonym analysis, or inferred meaning. Examples:

```
> family moved west
Statement verified. Original source: In early March 1830, fearing a milk sickness outbreak along the Ohio River, the Lincoln family moved west to Illinois, a non-slaveholding state.

> lincoln moved west
Statement could not be verified.

> family went west
```

```
Statement could not be verified.
```

We do not capture negations of meaning. Cases like this are why we display the source of "confirmation," so the user can make sure for himself the result makes sense:

```
> he was president
Statement verified. Original source: He was the first president from
the Republican Party.

> he was not president
Statement verified. Original source: He was the first president from
the Republican Party.
```

Evaluation: Our system is fairly narrow in terms of what it can do. It can verify if a statement is a restructuring or simplification of one in the source document. The biggest downside is that it really only supports restructuring, not *rewording*. Word choice on the input side must match word choice on the source side. It can handle different tenses/pluralities (via stemming), and different sentence structures / word orderings.

As far as a comparison to BM25 / skip bigrams, this system fulfills a fundamentally different purpose from our previous system. The information retrieval system looked for relevant documents to a general query. This natural language processing system seeks to give answers to specific queries. Each is relatively competent at its particular goal, and they could even work well together (similar to how Google will now provide answers to specific questions, as well as relevant documents to any arbitrary query).

The one way in which this could be seen as an improvement over the IR system is that it retrieves much more specific information. The IR system returns relevant documents, while in this NLP system if you specify a statement which actually finds a match, you get results at the granularity of a sentence.