Problems to deal with / think about:

Corpus processing:

1. Removing <code> tags and contents ? …
   1. Remove code tags when they refer to a long code block, but retain when it is simply identifying a function, package, etc (actually not a problem if my model ils trained only on titles, as these don’t have any html tags/formatting)
2. E.g, Title: Difference between Math.Floor() and Math.Truncate()
   1. Content: What is the difference between and in .NET?
   2. Notice in the content that the function call has been removed
3. Tokenizing while retaining periods when they denote a function or a name
4. Tokenizing special names (“.net”, “pandas.dataframe.iterparse”)
5. Tokenizing numbers (python2.7)
6. Misspellings – run a spell checker?
7. In the original code, there's a word2phrase preprocessing step you can do - essentially extracting most common bigrams from text. "language model" becomes "language\_model" training is performed as usual. You'll now have a vector for the phrase "language model"
8. Lemma-tizing / stemming words … so that suggestions aren’t simply other versions of word

Model predictions:

* Works best for verbs
* Not all words deserve substitutions – need to run each word against an algorithm that decides whether or not it should provide a substitution?
* Recurrent neural network (records an embedded layer of meaning)
* perceptron (standard NN)
* parts of speech

Future directions:

* Incorporate docs from packages

[How to find all questions that were duplicates of another question?](https://meta.stackexchange.com/questions/137503/how-to-find-all-questions-that-were-duplicates-of-another-question)

[What is best way to find duplicate questions? [duplicate]](https://meta.stackexchange.com/questions/160719/what-is-best-way-to-find-duplicate-questions)

[See how many duplicates of a post?](https://meta.stackexchange.com/questions/217415/see-how-many-duplicates-of-a-post)

https://meta.stackexchange.com/questions/linked/[question id]

For example: <https://meta.stackexchange.com/questions/linked/194476>

url:"http://meta.stackexchange.com/questions/7237/\*" closed:1

<http://data.stackexchange.com/stackoverflow/query/162748/duplicates-for-a-given-post?OriginalPostId=8028957>

select PostId as [Post Link]

from PostLinks

where RelatedPostId=##OriginalPostId##

and LinkTypeId=3 -- duplicate

<https://stats.stackexchange.com/questions/208213/logistic-regression-with-binomial-data-in-python>