# 需要做的事情

1. 从网站<https://www.sec.gov/edgar/searchedgar/companysearch.html> 上面下载对应的CIK number的 所有年份的10-k 和10-KSB report。(从1994年开始)

文件包括 10-k ，10-K/A, 10K405, 10-K405/A, 10-KSB, 10-KSB/A, 10-KSB40 and 10-KSB40/A

1. We prefer 10-K and 10-K405 over 10-KSB and 10-KSB40. We prefer any of those to the appendix filings. Only the most preferred filing is kept for each firm and fiscal year.
2. We delete all filings with less than 2000 tokens.
3. We further drop all observations prior to fiscal year 1995, as filing via EDGAR only became mandatory in May 1996.
4. For each filing, we select the Form 10-K document only and exclude all exhibits, PDFs, Excel files and images attached to the filing.
5. Next, we remove all XBRL and ASCII data from the document.
6. Parsing the document into a tree structure, allows us to identify specific HTML elements.
7. We select and delete all tables with more than 15% numeric characters, because such tables do most likely not contain actual text.
8. Using the **Jericho Java library**, we extract all remaining text from the reduced Form 10-K.
9. If a Form 10-K document is not stored in HTML format, which is the case for older filings, we still run this procedure, but won’t be able to detect and remove tables.
10. The raw text data is split into sentences using the **Stanford Natural Language Processing Toolkit library** presented in Manning et al. (2014).
11. We store sentences when they contain at least one word from our uncertainty word list.
12. For each document we go through all sentences and construct word counts using the overall uncertainty, risk and ambiguity word lists.
13. We also calculate the highest cosine similarity of each sentence with any sentence of the firm’s annual report from the previous year and identify whether the sentence contains any negation terms.
14. As measures of document length, we count the number of sentences and words that occur in the complete Loughran and McDonald (2011) dictionary.

# 2. Parsing Procedure for the 10-K Sample

We first download all 10-K and 10-K405 documents identified in the quarterly master index files appearing on the EDGAR website for the period 1994 to 2008. Each complete text filing is read into a single string variable and parsed using the following sequence:

* Remove graphics (ASCII encoded graphics)
* Identify self-reported SIC code on the first page of the filing. If the SIC code does not appear in the 10-K, we programmatically go to the general web page for the firm on the EDGAR site to see if a SIC code is reported. If no SIC code is found, the industry is classified as “Other.”
* Remove SEC header—we remove the standard first page of the filing appearing between the HTML or tags.
* Re-encode characters—translates “encoded” characters such as &NBSP (blank space) or & (&) back to their original ACSII form.
* Remove exhibits—removes all text appearing within “EX” HTML tagged document segments.
* Remove tables—we remove all text appearing within <TABLE> HTML tags, where more than 25% of the nonblank characters are numbers. (Note that this filter is important since some filers embed all of their text within TABLE tags.)
* Remove HTML—the quantity of HTML increased substantially in the late 1990s. Some 10-Ks include more HTML than actual content.
* Parse into tokens—we use a regular expression (regex) to parse the remaining string variable into all collections of two or more alphabetic characters. (Hyphens are also allowed in the character collections.) We first replace all hyphens followed by a line-feed with a hyphen so that the word boundary regex works correctly.
* Create word counts—at this point we have a collection of alphabetic characters (tokens), which we then look up in our master dictionary. This parsing process also accounts for hyphenation. We keep a word count for all words in the master dictionary for each document. This allows us to subsequently go back and create word counts based on the various tonal word lists using the document dictionaries.

# 3. Word list

**Ambiguity words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ABEYANCE | DOUBTED | PRESUMPTIONS | UNCERTAINTY | UNPROVED |
| ABEYANCES | DOUBTFUL | REINTERPRET | UNCLEAR | UNPROVEN |
| AMBIGUITIES | DOUBTS | REINTERPRETATION | UNCONFIRMED | UNQUANTIFIABLE |
| AMBIGUITY | HIDDEN | REINTERPRETATION | UNDECIDED | UNQUANTIFIED |
| AMBIGUOUS | IMPRECISE | REINTERPRETED | UNDEFINED | UNRECONCILED |
| ANOMALIES | IMPRECISION | REINTERPRETING | UNDESIGNATED | UNSEASONABLE |
| ANOMALOUS | IMPRECISIONS | REINTERPRETS | UNDETECTABLE | UNSEASONABLY |
| ANOMALOUSLY | INCOMPLETENESS | REVISE | UNDETERMINABLE | UNSPECIFIC |
| ANOMALY | INDEFINITE | REVISED | UNDETERMINED | UNSPECIFIED |
| ARBITRARILY | INDEFINITENESS | RUMORS | UNDOCUMENTED | UNTESTED |
| ARBITRARINESS | INDETERMINABLE | SELDOM | UNEXPECTED | UNUSUAL |
| ARBITRARY | INDETERMINATE | SELDOMLY | UNEXPECTEDLY | UNUSUALLY |
| BELIEVE | INEXACT | SPECULATE | UNFAMILIAR | VAGARIES |
| BELIEVED | INEXACTNESS | SPECULATED | UNFAMILIARITY | VAGUE |
| BELIEVES | MAYBE | SPECULATES | UNFORECASTED | VAGUELY |
| BELIEVING | MIGHT | SPECULATING | UNFORSEEN | VAGUENESS |
| CAUTIOUS | NONASSESSABLE | SPECULATION | UNGUARANTEED | VAGUENESSES |
| CAUTIOUSLY | PERHAPS | SPECULATIONS | UNIDENTIFIABLE | VAGUER |
| CAUTIOUSNESS | PRECAUTION | SPECULATIVE | UNIDENTIFIED | VAGUEST |
| CONCEIVABLE | PRECAUTIONARY | SPECULATIVELY | UNKNOWN | |
| CONCEIVABLY | PRECAUTIONS | SPORADIC | UNKNOWNS | |
| CONFUSES | PRESUMABLY | SPORADICALLY | UNOBSERVABLE | |
| CONFUSING | PRESUME | SUDDEN | UNPLANNED | |
| CONFUSINGLY | PRESUMED | SUDDENLY | UNPREDICTABILITY | |
| CONFUSION | PRESUMES | UNCERTAIN | UNPREDICTABLE | |
| COULD | PRESUMING | UNCERTAINLY | UNPREDICTABLY | |
| DOUBT | PRESUMPTION | UNCERTAINTIES | UNPREDICTED | |

**Risk words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ANTICIPATE | DEVIATING | PREDICT | RANDOMIZE | VARIANCE |
| ANTICIPATED | DEVIATION | PREDICTABILITY | RANDOMIZED | VARIANCES |
| ANTICIPATES | DEVIATIONS | PREDICTED | RANDOMIZES | VARIATION |
| ANTICIPATING | EXPOSURE | PREDICTING | RANDOMIZING | VARIATIONS |
| ANTICIPATION | EXPOSURES | PREDICTION | RANDOMLY | VARIED |
| ANTICIPATIONS | FLUCTUATE | PREDICTIONS | RANDOMNESS | VARIES |
| APPROXIMATE | FLUCTUATED | PREDICTIVE | RISKIER | VARY |
| APPROXIMATED | FLUCTUATES | PREDICTOR | RISKIEST | VOLATILE |
| APPROXIMATELY | FLUCTUATING | PREDICTORS | RISKINESS | VOLATILITIES |
| APPROXIMATES | FLUCTUATION | PREDICTS | RISKY | VOLATILITY |
| APPROXIMATING | FLUCTUATIONS | PROBABILISTIC | SOMETIME | |
| APPROXIMATION | IMPROBABILITY | PROBABILITIES | SOMETIMES | |
| APPROXIMATIONS | IMPROBABLE | PROBABILITY | VARIABILITY | |
| DEVIATE | LIKELIHOOD | PROBABLE | VARIABLE | |
| DEVIATED | OCCASIONALLY | PROBABLY | VARIABLES | |
| DEVIATES | ORDINARILY | RANDOM | VARIABLY | |

**Unclassified uncertainty words**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ALMOST | CONTINGENCY | INDEFINITELY | RECALCULATE | SOMEWHERE |
| ALTERATION | CONTINGENT | INSTABILITIES | RECALCULATED | SUGGEST |
| ALTERATIONS | CONTINGENTLY | INSTABILITY | RECALCULATES | SUGGESTED |
| APPARENT | CONTINGENTS | INTANGIBLE | RECALCULATING | SUGGESTING |
| APPARENTLY | CROSSROAD | INTANGIBLES | RECALCULATION | SUGGESTS |
| APPEAR | CROSSROADS | MAY | RECALCULATIONS | SUSCEPTIBILITY |
| APPEARED | DEPEND | NEARLY | RECONSIDER | TENDING |
| APPEARING | DEPENDED | PENDING | RECONSIDERED | TENTATIVE |
| APPEARS | DEPENDENCE | POSSIBILITIES | RECONSIDERING | TENTATIVELY |
| ASSUME | DEPENDENCIES | POSSIBILITY | RECONSIDERS | TURBULENCE |
| ASSUMED | DEPENDENCY | POSSIBLE | REEXAMINATION | UNHEDGED |
| ASSUMES | DEPENDENT | POSSIBLY | REEXAMINE | UNSETTLED |
| ASSUMING | DEPENDING | PRELIMINARILY | REEXAMINING | UNWRITTEN |
| ASSUMPTION | DEPENDS | PRELIMINARY | RISK | VARIANT |
| ASSUMPTIONS | DESTABILIZING | REASSESS | RISKED | VARIANTS |
| CLARIFICATION | DIFFER | REASSESSED | RISKING | VARYING |
| CLARIFICATIONS | DIFFERED | REASSESSES | RISKS |  |
| CONDITIONAL | DIFFERING | REASSESSING | ROUGHLY |  |
| CONDITIONALLY | DIFFERS | REASSESSMENT | SEEMS |  |
| CONTINGENCIES | HINGES | REASSESSMENTS | SOMEWHAT | |

# 4. 一共7个需要的变量

1. Uncertainty的比重（单词）

这一共的298个uncertainty(3个带括号tab里面的所有单词) 的单词，占当年的10-K的单词总量的百分比 (分母是所有的出现的在LoughranMcDonald\_MasterDictionary\_2016.xlsx的单词总量，重复计算)

1. Ambiguity 的比重 （单词）

这一共的127个Ambiguity的单词，占当年的10-K的单词总量的百分比 (分母是所有的出现的在LoughranMcDonald\_MasterDictionary\_2016.xlsx的单词总量，重复计算)

1. Risk 的比重 （单词）

这一共的74个 Risk的单词，占当年的10-K的单词总量的百分比 (分母是所有的出现的在LoughranMcDonald\_MasterDictionary\_2016.xlsx的单词总量，重复计算)

1. Ambiguity 的比重 （句子）

所有包含至少一个Ambiguity的单词的句子，占当年的10-K的所有句子的百分比。（如果一个句子既包含ambiguity 又 包含risk等， 就算两次）

1. Risk 的比重 （句子）

所有包含至少一个Risk的单词的句子，占当年的10-K的所有句子的百分比。（如果一个句子既包含ambiguity 又 包含risk等， 就算两次）

1. Negative (单词)

这一共的2355个negative的单词，占当年的10-K的单词总量的百分比 (分母是所有的出现的在LoughranMcDonald\_MasterDictionary\_2016.xlsx的单词总量，重复计算)

1. Positive (单词)

这一共的354个positive的单词，占当年的10-K的单词总量的百分比 (分母是所有的出现的在LoughranMcDonald\_MasterDictionary\_2016.xlsx的单词总量，重复计算)

Use Wharton Data Services CIK ﬁle to link SEC CIK numbers to the CRSP PERMNOs