Analyse of insulin article

May 3, 2017

In [1]:
```
import nltk
from collections import Counter
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.stem.wordnet import WordNetLemmatizer
import pandas
```

In [2]:
```
lines = "Insulin (from the Latin, insula meaning island) is a peptide hormone produced b
        "islets, and by the Brockmann body in some teleost fish. It has important effects
        "carbohydrates, fats and protein by promoting the absorption of, especially, gluc
        "liver and skeletal muscle cells. In these tissues the absorbed glucose is conver
        "glycogenesis or fats (triglycerides) via lipogenesis, or, in the case of the liv
        "(and excretion into the blood) by the liver is strongly inhibited by high concen
        "blood. Circulating insulin also affects the synthesis of proteins in a wide vari
        "in the blood it is therefore an anabolic hormone, promoting the conversion of sm
        " large molecules inside the cells. Low insulin levels in the blood have the oppo
        "catabolism. The pancreatic beta cells are known to be sensitive to the " \
        "glucose concentration in the blood. When the blood glucose levels are high they
        "when the levels are low they cease their secretion of this hormone into the gene
        " alpha cells, probably by taking their cues from the beta cells, secrete glucago
        "opposite manner: high secretion rates when the blood glucose concentrations are
        " when the glucose levels are high. High glucagon concentrations in the blood pla
        "liver to release glucose into the blood by glycogenolysis and gluconeogenesis, t
        "on the blood glucose level to that produced by high insulin concentrations. The
        "into the blood in response to the blood glucose concentration is the primary mec
        " the glucose levels in the extracellular fluids within very narrow limits at res
        " exercise and starvation. When the pancreatic beta cells are destroyed by an aut
        "insulin can no longer be synthesized or be secreted into the blood. This results
        "which is characterized by very high blood sugar levels, and generalized body was
        "This can only be corrected by injecting the hormone, either directly into the bl
        "and confused or comatosed, or subcutaneously for routine maintenance therapy, wh
        "rest of the person's life. The exact details of how much insulin needs to be inj
        " has to be adjusted according to the patient's daily routine of meals and exerci
        "physiological secretion of insulin as closely as is practically possible."
```

In [3]:
```
noun_tags = ['NN', 'NNP', 'NNS', 'NNPS']
adj_tags = ['JJ', 'JJR', 'JJS']
```

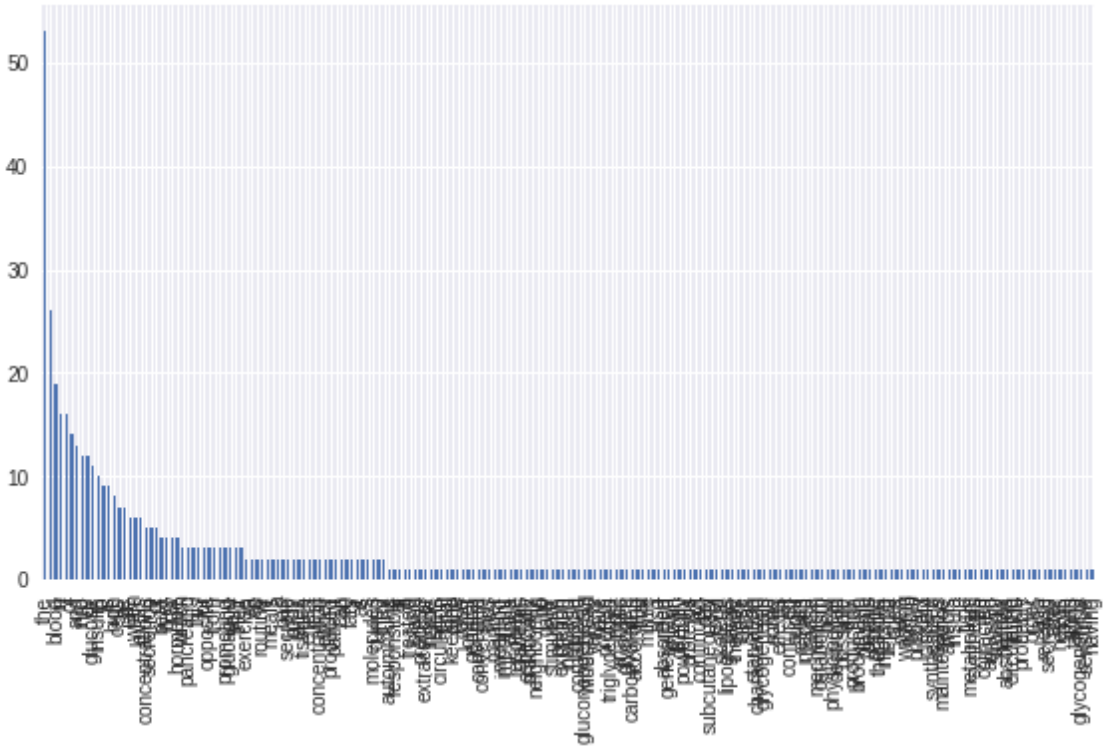
```
        adv_tags = ['RB', 'RBR', 'RBS']
        verb_tags = ['VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ']
```

```
In [20]: def get(tags):
             results = []
             for sentence in nltk.sent_tokenize(lines):
                 for word,pos in nltk.pos_tag(nltk.word_tokenize(str(sentence))):
                     if len(tags) > 0:
                         if pos  in tags:
                             results.append(word.lower())
                     else:
                         results.append(word.lower())
             return results

         def show(data, rot=90):
             print(str(len(data)) + " instances - " + str(len(set(data))) + " distincts.")
             if len(data) > 1:
                 c = Counter(data)
                 df = pandas.DataFrame.from_dict(c, 'index')
                 df = df.sort_values([0], ascending=False)
                 df.plot(kind='bar', legend=False, rot = rot)
                 plt.tight_layout()
                 plt.show()
```

```
In [21]: show(get([]))
```

```
515 instances - 199 distincts.
```

```
<matplotlib.figure.Figure at 0x7fe0529d58d0>
```



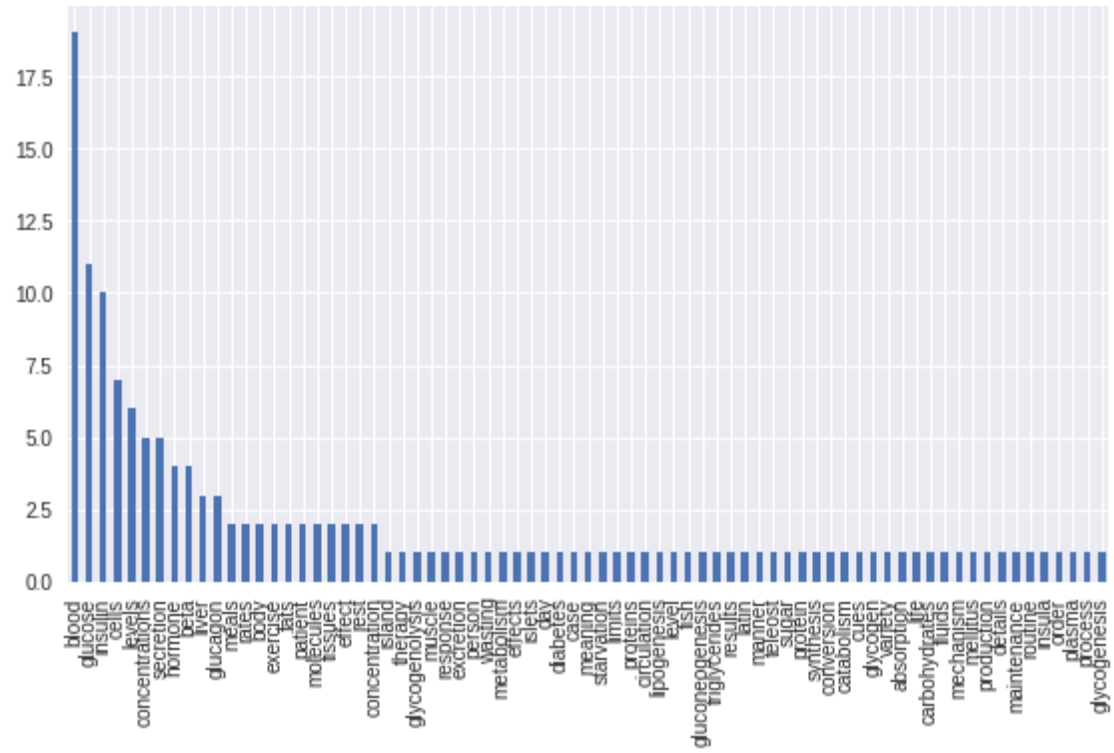
In [6]:
```
nouns = get(noun_tags)
nouns.sort()
print nouns
```

```
['absorbed', 'absorption', 'alpha', 'beta', 'beta', 'beta', 'beta', 'blood', 'blood', 'blood', '
```

In [7]:
```
# Correcting some of the errors (brockmann-, absorded-, glucose+ )

nouns = [ 'absorption', 'beta', 'beta', 'beta', 'beta', 'blood', 'blood', 'blood', 'bloo
          'blood', 'blood', 'blood', 'blood', 'blood', 'blood', 'blood', 'blood', 'blood'
          'blood', 'blood', 'body', 'body', 'carbohydrates', 'case', 'catabolism', 'cells
          'cells', 'cells', 'cells', 'cells', 'circulation', 'concentration', 'concentrat
          'concentrations', 'concentrations', 'concentrations', 'concentrations', 'conver
          'diabetes', 'effect', 'effect', 'effects', 'excretion', 'exercise', 'exercise',
          'fluids', 'glucagon', 'glucagon', 'glucagon', 'gluconeogenesis',  'glucose', 'g
          'glucose','glucose', 'glucose', 'glucose', 'glucose', 'glucose', 'glucose', 'gl
          'glycogenesis', 'glycogenolysis', 'hormone', 'hormone', 'hormone', 'hormone', '
          'insulin', 'insulin', 'insulin', 'insulin', 'insulin', 'insulin', 'insulin', 'i
          'latin', 'level', 'levels', 'levels', 'levels', 'levels', 'levels', 'levels', '
          'liver', 'liver', 'liver', 'maintenance', 'manner', 'meals', 'meals', 'meaning'
          'metabolism', 'molecules', 'molecules', 'muscle', 'order', 'patient', 'patient'
          'process', 'production', 'protein', 'proteins', 'rates', 'rates', 'response', '
```

```
                'routine', 'secretion', 'secretion', 'secretion', 'secretion', 'secretion', 'st
                'synthesis', 'teleost', 'therapy', 'tissues', 'tissues', 'triglycerides', 'vari

        nouns.sort()
        show(nouns)

        nouns = list(set(nouns))
        nouns.sort()
        print nouns
```

```
150 instances - 73 distincts.
```




```
['absorption', 'beta', 'blood', 'body', 'carbohydrates', 'case', 'catabolism', 'cells', 'circula
```

```
In [8]: done = ['tissues', 'cells', 'carbohydrates', 'protein', 'proteins', 'insulin',
                'person', 'patient', 'fats', 'fluids', 'rate', 'concentration',
                'levels', 'level', 'rates', 'cues', 'details', 'fish', 'blood',
                'glucose', 'sugar', 'glycogen', 'glucagon', 'hormone', 'liver',
                'process', 'secretion']
        not_cons = ['beta', 'insula', 'latin', 'teleost']
```

```
        remainings = [n for n in nouns if n not in done and n not in not_cons]
        print str(len(remainings))
        print remainings
```

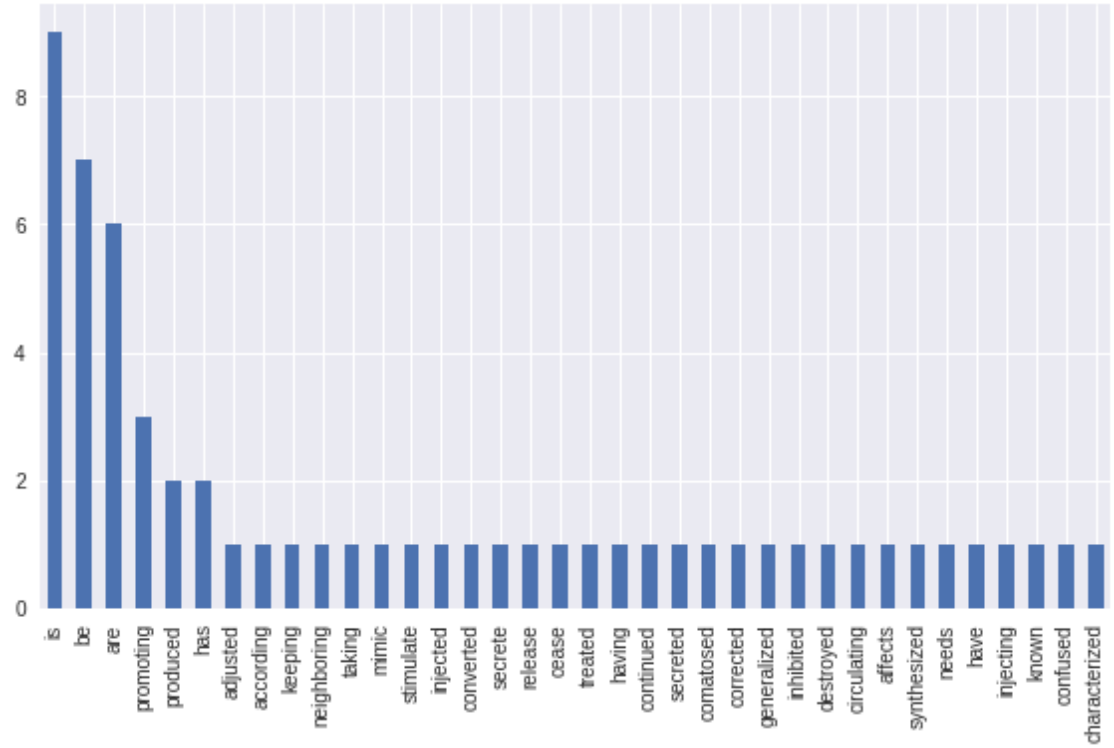
```
43
['absorption', 'body', 'case', 'catabolism', 'circulation', 'concentrations', 'conversion', 'day
```

In [9]:
```
        verbs = get(verb_tags)
        verbs.sort()
        print verbs
```

```
['according', 'adjusted', 'affects', 'are', 'are', 'are', 'are', 'are', 'are', 'be', 'be', 'be',
```

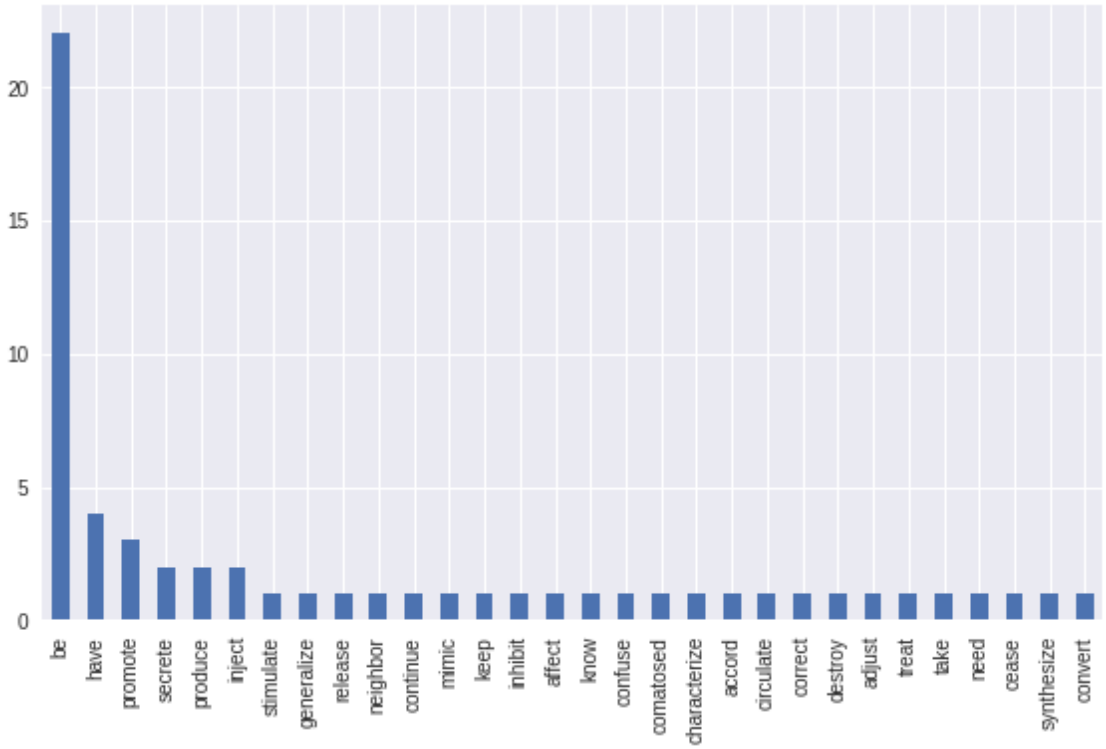
In [10]:
```
        verbs = ['according', 'adjusted', 'affects', 'are', 'are', 'are', 'are', 'are', 'are',
                 'be', 'be', 'be', 'be', 'be', 'be', 'be', 'cease', 'characterized', 'circulati
                 'comatosed', 'confused', 'continued', 'converted', 'corrected', 'destroyed', '
                 'has', 'has', 'have', 'having', 'inhibited', 'injected', 'injecting',
                 'is', 'is', 'is', 'is', 'is', 'is', 'is', 'is', 'is', 'keeping', 'known', 'mim
                 'produced', 'produced', 'promoting', 'promoting', 'promoting', 'release', 'sec
                 'stimulate', 'synthesized', 'taking', 'treated']

        show(verbs)
```

```
59 instances - 36 distincts.
```



```
In [23]: verbs_inf = [str(WordNetLemmatizer().lemmatize(v, 'v')) for v in verbs]
         show(verbs_inf)

         verbs_inf = list(set(verbs_inf))
         verbs_inf.sort()
         print verbs_inf
```

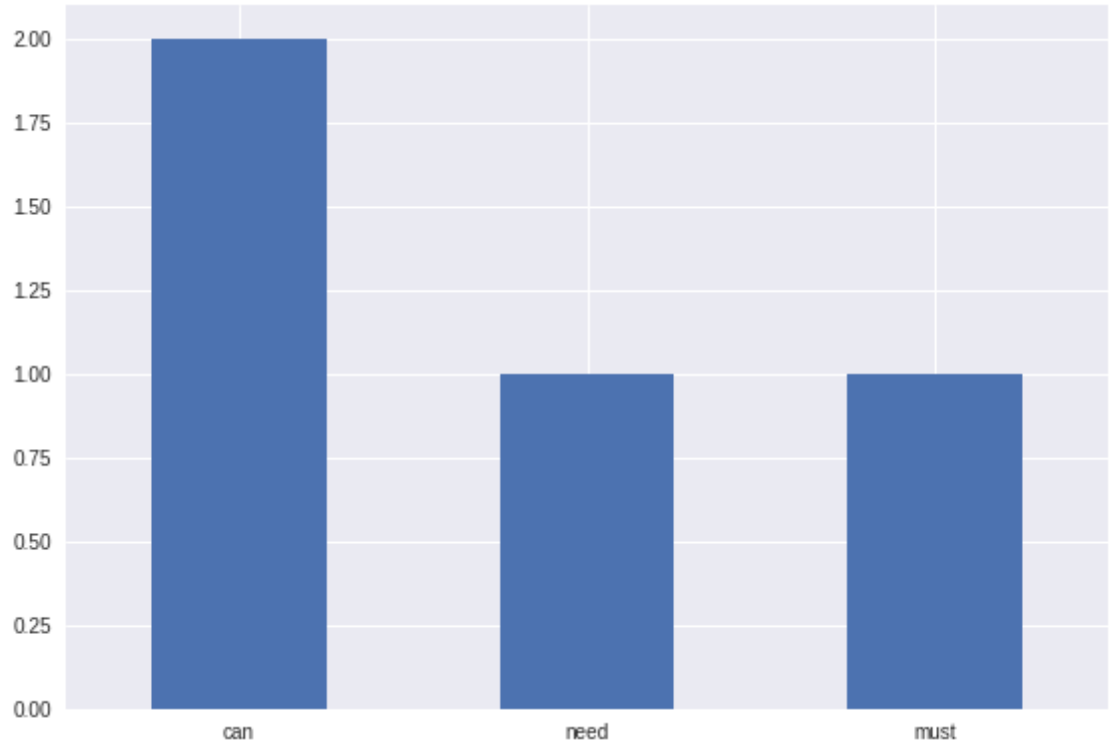
```
59 instances - 30 distincts.


<matplotlib.figure.Figure at 0x7fe052e89ed0>
```




```
['accord', 'adjust', 'affect', 'be', 'cease', 'characterize', 'circulate', 'comatosed', 'confuse
```

```
In [12]: aux = get(['MD'])
         aux.append('need')
         show(aux, 0)
```

```
4 instances - 3 distincts.
```

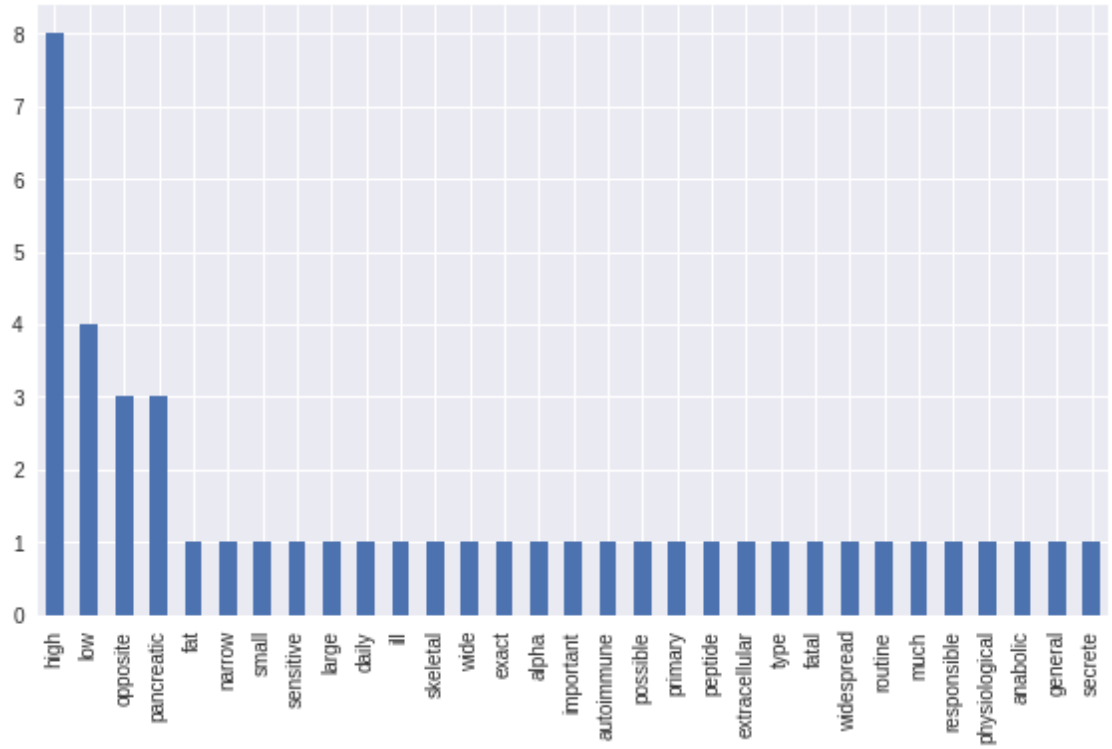



```
In [13]: adjs = get(adj_tags)
         adjs.sort()
         print adjs
```

```
['anabolic', 'autoimmune', 'daily', 'exact', 'extracellular', 'fat', 'fatal', 'general', 'glucos
```

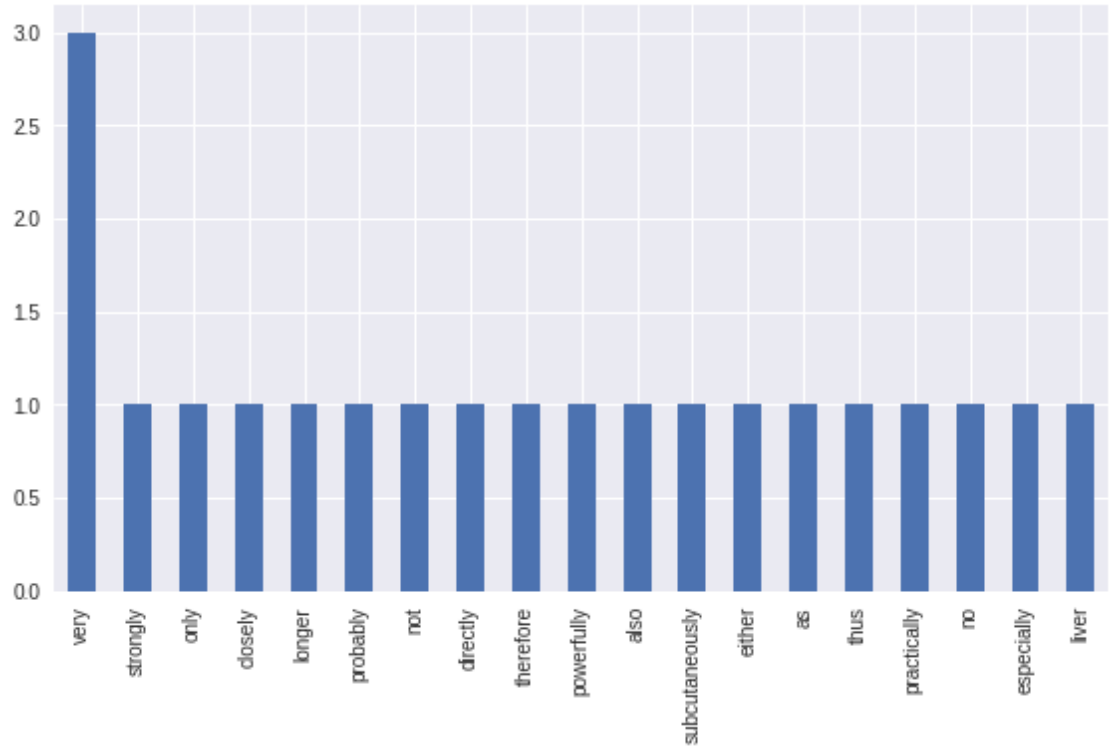
```
In [14]: adjs = ['alpha', 'anabolic', 'autoimmune', 'daily', 'exact', 'extracellular', 'fat', 'f
                 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'high', 'ill', 'importa
                 'low', 'low', 'low', 'low', 'much', 'narrow', 'opposite', 'opposite', 'opposite
                 'pancreatic', 'pancreatic', 'pancreatic', 'peptide', 'physiological', 'possible
                 'routine', 'secrete', 'sensitive', 'skeletal', 'small', 'type', 'wide', 'widesp

         show(adjs)
```

```
45 instances - 31 distincts.
```



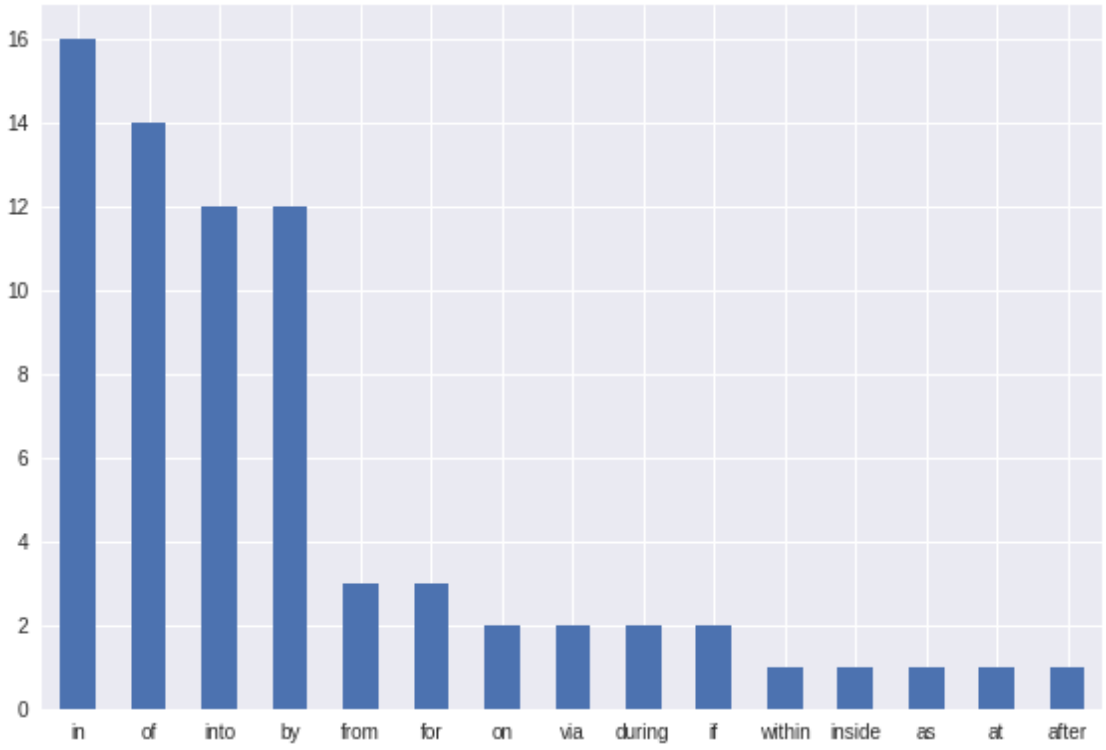
Glucose appears here as adjectives due to his involvment in some compound nouns constructions. So it has been deleted.

```
In [15]: advs = get(adv_tags)
         show(advs)

21 instances - 19 distincts.
```




```
In [16]: cin = get(['IN'])
         show(cin, 0)
```

```
73 instances - 15 distincts.
```




```
In [17]: show(['EX'])
```

```
1 instances - 1 distincts.
```

```
In [18]: cc = get(['CC'])
         show(cc, 0)
```

```
18 instances - 2 distincts.
```