Arbeitsprotokoll

Die erste Frage, die sich stellt, ist: welche Dokumente sollen ausgewertet werden. Die einführenden Kommentare zum Zinsentscheid (Introductory Statements) oder die etwas längeren Sitzungsprotokolle (Monetary Accounts).

Die IS haben den Vorteil, dass sie seit Einführung des Euros als buchhalterische Währung im Jahr 1998 verfasst und in mehreren Sprachen publiziert werden. Sie sind nach einem sich über die Zeit nur wenig ändernden Schema verfasst, was für strukturierte Auswertungen von Vorteil sein kann. Der Nachteil ist, dass die Texte fast identisch aufgebaut sind und sich nur wenige Passagen ändern, die allgemein bekannt sind. Schon heute werden die Statements von den EZB-Watchern am Markt Wort für Wort seziert. Da die entscheidenden Passagen kurz sind, braucht es keine speziellen Programmierkenntnisse.

Mehr versteckte Hinweise auf die Stimmungslage innerhalb der Zentralbank versprechen die längeren Protokolle. Nur wenige Analysten nehmen sich die Zeit, die Dokumente zu lessen. Die Accounts werden aber erst seit vier Jahren publiziert, wodurch Aussagen auf die kurze Frist beschränkt sind.

Ich entscheide mich für ein stufenweises Vorgehen. Zuerst warden die Protokolle ausgewertet. Dann zu Vervollständigung und als Cross-Check die Introductionary Statements. Sind die Accounts für Fragestellungen mit Aktualitätsbezug vielversprechend, könnten die IS interessante Ergebnisse zu längerfristigen Entwicklungen geben.

Auswertung der Sitzungprotokolle (Monetary accounts)

Die Deutsche Bundesbank (Buba) hat die 33 Protokolle auf ihrer Website auf Deutsch als PDF verfügbar gemacht. PDF können mit textract relativ einfach eingelesen werden. Das oft zeitaufwändige „Scrapen“ der Webseiten entfällt.

Wenn man aber die 33 PDF nicht von Hand einzeln herunterladen will und das im CAS-Kurs vermittelte Wissen anwenden möchte, kann man auf das Scrapen nicht ganz verzichten.

Datenbeschaffung (vgl. ezbprotokolle_datenbeschaffung.ipynb)

Mit Hilfe von Requests und BeautifulSoup lassen sich die URLs zu den PDF relativ schnell sammeln und in eine Liste packen. Mit dieser URL-Liste werden dann die PDF automatisch heruntergeladen.

Die Seite der Buba ermöglicht es, bis zu 50 Dokumente anzeigen zu lassen. Das ist die URL, die ich ansteuern möchte:
https://www.bundesbank.de/action/de/737380/bbksearch?state=H4sIAAAAAAAAAG2QQWrDQAxFr1K0LF4421kGYiiE1pBcQB4r9lBlxtHICxNy92ggNi1kpy89fb50h9tMsoCDT6jAjxgj8VcPLs7MFfTJn5eJVjnhQN_zFVxdwQU9aQZ3f1gdWEleQtMwMDVlvF9-Jg0prvtj0NyStGYDblf_a5xUQhwsyK6G4ihZj6Rmuy4jTyN2pAaHiMX2ELFjsrAX5EzviFYoE5PXDVWZjexRqZF03e40fU5_1SmxNIF4e0VOopau635LVZCPnrKHxxMTzqlXQwEAAA&hitsPerPageString=50&sort=bbksortdate+desc

Mit Hilfe des Developer-Tools werden die Links zu den PDF herausgezogen. Sie sind mit dem Tag “a” und “href” gekennzechnet. Die Seite enhält enthält über 800 Links. Man muss die Liste zurechtschneiden. Danach erhält man eine Liste mit den 33 Urls zu den PDF.
Damit die PDF herunterladen und speichern.

Textanalyse (ezbprotokolle_textanalyse.ipynb)

PDF mit mit Textract einlesen. Zuerst für als Test für den Einzelfall, dann mit einer Schlaufe für die ganze Liste.

Das erfodert Rechen-Power und dauert. Es empfiehlt sich, eine Progressbar einzubauen.

Aus dem gewonnen Dictionary mit den Keys “Filename” und “Text” wird ein Pandas-Dataframe erstellt

Danach wird das Dataframe so bearbeitet, dass das Datum zum Index wird. Resultat ist ein simples Dataframe mit dem Datum als Index und den Spalten “Filename” und “Text”.

Damit lassen sich einfache Abfragen durchführen, etwa wie oft bestimmte Begriffe in den 33 Dokumenten vorkommen.

Das geht mit der einfachen count-Funktion oder der Regex-Funktion “Find all”. In diesem Fall hab ich mich für die Regex-Funktion entschieden.

Die Resultate werden in einfachen Liniengrafiken dargestellt, die den Verlauf der Häufigkeit der Begriffe abbilden.

Inspiriert von den Gespräche mit den Briefing-Personen und aufgrund eigener Überlegungen werden die Protokolle auf verschieden Begriffe oder Wortteile geprüft. Interessante Resultate und Muster werden gespeichert, trendlose Kurven ohne Aussagen verworfen.

Suchbegriffe sind:

Inflation, Teuerung, Inflationserwartungen, Kerninflation und verwandte Begriffe (dazu zählen etwa “zugrunde liegende Inflation” oder “ohne Energie und Nahrungsmittel”), Risiko/Risiken, Unsicherheit, Deflation, Abwärtsrisiken, Verlangsamung, schwächer, Wechselkurs, Auf- und Abwertung, Schwellenländer, China, akkommodierend bzw. Akkommodierung (so bezeichnet die EZB eine unterstützende Geldpolitik)

Die wichtigsten Resultate:

- Die Kerninfaltion war bis vor wenigen Monaten ein grosses Thema, mit der Abkühlung der Konjunktur und dem Rückgang der Inflationserwartungen jedoch nicht mehr.
- Abwärtsrisiken und Unsicherheiten haben zugenommen, sie sind auf nahe Rekordniveaus
- Begriffe, die die Abkühlung thematisieren, haben Hochkonjunktur
- Die Aufwertung des Euro war während der Eurostärke 2017/18 ein Thema

- Schwellenländer warden häufiger erwähnte, wegen ihrer Anfälligkeit vor allem als Risikofaktor

Die einzelnen Textdokumente lassen sich auch anderweitig analysieren und bearbeiten: Ein Beispiel sind Histogramme mit den häufigsten Wörtern oder Wordclouds ohne Stopp-Wörter. Das erfordert viele Instrumente des NLTK-Toolkit. Die Ergebnisse sind aber nicht so spektakulär und werden deshalb nicht mehr weiterverfolgt.

Interessant hingegen wäre eine Sentiment-Analyse: Wie positiv oder negative sind die Protokolle formuliert? Das Problem allerding ist, dass im Web sehr wenig über Sentiment-Analyse von deutschen Texten zu finden ist. Mit Hilfe des Emoticons der Universität Leipzig (SentiWS-v2.0) könnte man einen eigenen Klassifier erstellen und so eine Sentiment-Tool bauen. Der Aufwand ist aber enorm. In den Unterlagen stosse ich schliesslich auf die Python-Paket TextblobDE. Es erlaubt, mit wenig Code, die Polarität der Protokolle zu bestimmen. Auch wenn TexblobDE Mühe hat, die nüchterne Sprache der Ökonomen zu bewerten (schwächer als erwartet bekommt zum Beispiel einen Polaritätswert von 0!) sind die Ergebnisse in der Summe plausibel.

Die Protokoller werden im Zuge der Erholung 2017 immer positiver. Ab 2018 fällt das Polaritätsmass aber wieder steil ab.

Als Teil einer Geschichte zur EZB-Geldpolitik konnten die Ergebnisse am 26. Februar in der Finanz & Wirtschaft publiziert werden.

Die bereinigten Skripts der Analyse sind als ezbprotokolle_datenbeschaffung.ipynb und ezbprotokolle_Textanalyse.ipynb in der Github-Repository “project_ecb_text” abgelegt.

Auswertung der Introductory Statements (vgl. ecb_is.ipynb)

Beschaffung der Daten

Auf der Webseite der EZB sind die IS seit den Anfängen der Zentralbank 1998 aufgeschaltet. Dazu muss man unter Press conferences jedes Jahr anklicken, und es öffnet sich die Liste der IS des entsprechenden Jahres. Die URL zu den Texten könnte man über den Tag “href” scrapen. Ein Test und die Erfahrungen mit den Monetary Accounts haben aber gezeigt, dass das sehr aufwendig und mühsam ist, weil auf der Seite enorm viele andere Links aufgelistet sind (zB. Zu den Dokumenten in den verschiedenen Sprachen)

Deshalb entscheide ich mich für eine einfacherer Variante: Die URL lassen sich auch generieren, denn sie sind recht einfach aufgebaut. Hier ein Beispiel für das Statement zur PK vom 2. Dezember 1999:

https://www.ecb.europa.eu/press/pressconf/1999/html/is991202.en.html

Sie bestehen aus einer konstanten Basis-Url, dem Jahr und dem Datum.

Mit diesen Elementen lassen sich die URL erstellen: Achtung, ab dem 27. April 2017 ändert sich die Basisurl am Ende leicht.

Er hat die Form:

https://www.ecb.europa.eu/press/pressconf/2017/html/ecb.is170427.en.html

Zuerst wird eine Liste mit den sogenannten Basisurls erstellt, in denen sich nur das Jahr ändert: Die Elemente sehen so aus:

https://www.ecb.europa.eu/press/pressconf/2017/html/index.en.html, wobei sich das Jahr ändert.

Über diese Urls ziehen wir mit Hilfe von requests und BeautifulSoup das Datum der jeweiligen Pressekonferenz und erstellen daraus eine Liste mit den Daten.

Um diese List chronologisch zu ordnen, müssen die Element von einem String-Objekt in eine Datum geändert werden. Aus der Datumliste lässt sich über Umwege auch eine Liste erstellen, in denen das Datum die Form 990107 hat, wie in der URL.

Mit der fixen Basisurl, der Jahresliste und der Datumliste werden dann die URLs der Jahre 1999 bis 2018 zusammengebaut. Hinzugefügt wird die URL für die Sitzung vom Januar, die sich von den anderen deutlich unterscheidet.

https://www.ecb.europa.eu/press/pressconf/2019/html/ecb.is190124~cd3821f8f5.en.html

Die Liste der URLs wird dann gespeichert.

Texte extrahieren, bereinigen und in Dataframe laden

Nun müssen die Texte anhand der URL extrahiert werden. Ein Versuch, den mit p getagten Text zu scrapen scheitert. Zum Glück gibt es noch Smart Extraction.

Damit lässt sich der ganze Inhalt einer Seite ziehen. Mann muss aber die Bytes noch in Strings umwandeln und nur den oberen Teil der Seite bis zur Q&A Seite auswählen.

Man kann die Texte einfach nach einem bestimmten Character teilen (spliten). Für den Einzelfall funktioniert das bestens. Im aktuellen Statement wird die Q&A-Session mit drei Sternen vom Rest des Statements getrennt wird. Das Problem ist aber, dass in früheren Statements die drei Sterne fehlen. Gemäss Stichproben kommen neben den drei Sternen folgende Worte und Sätze zwischen IS und Q&A vor:

“We are now at your disposal” in verschiedenen Varianten. “at your disposal” scheint ein geeigneter Delimiter zu sein.

Oft beginnt die Q&A mit dem Hinweis “transcript” of the Q&A. Mit diesen beiden Delimiters werden alle 230 Texte am richtigen Ort getrennt.

Die Texte in ein Pandas-Dataframework. Es braucht noch ein paar Operationen, bis man ein DF mit dem Datum als Index und den Spalten URL und dem extrahierten Text zusammengestellt hat.

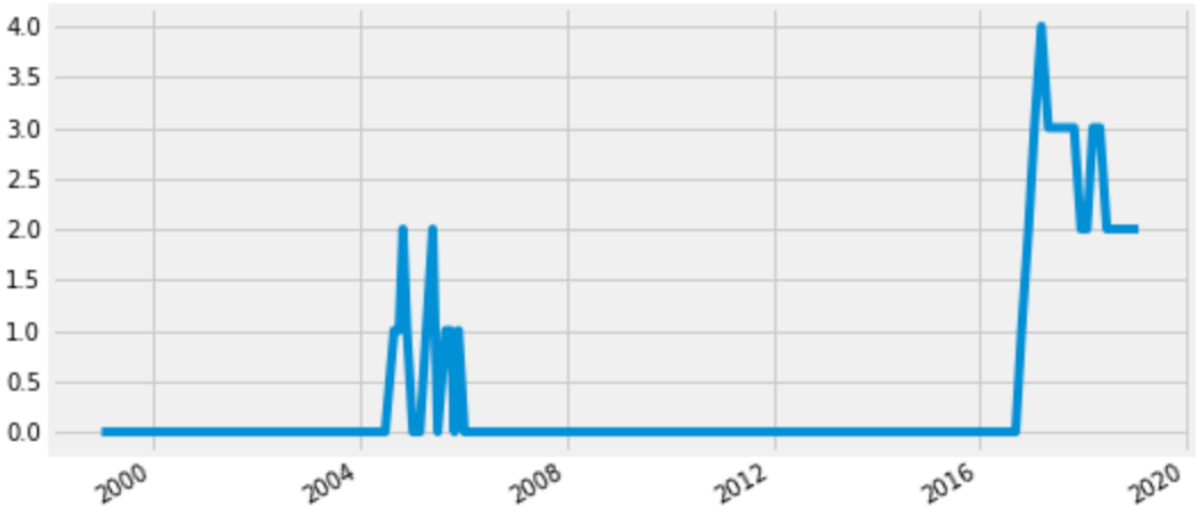
Textanalyse

Dann kann die Auswertung beginnen: Wie lang sind die Texte? Bei einigen wurde nu rein Teil des Textes extrahiert. Die Ausreisser werden herausgefiltert:

Wie bei den Monetary Accounts werden nun bestimmte Wortnennung gezählt, die geldpolitische relevant sind:

Wichtigste Ergebnisse:

Das Thema der Kerninflation (underlying inflation) taucht nur in zwei Phasen auf. Einmal im Boom 2005 und dann wieder ab 2017. Ähnlich wie bei den Monetary accounts geht die Zeitreihe Ende 2018 leicht zurück.



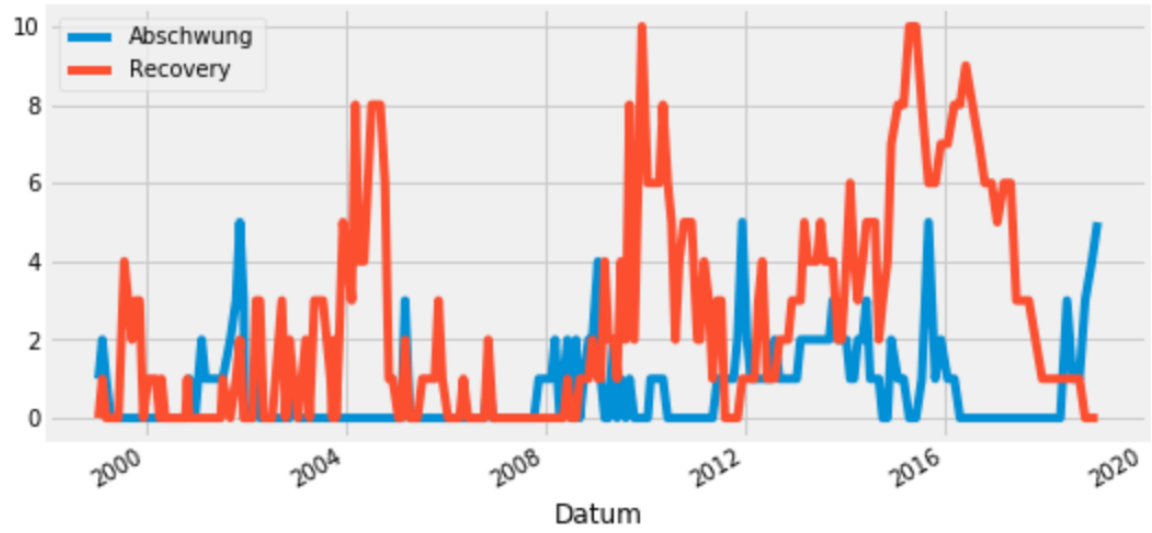
Die Wörter “accommodation” und “accommodative” kommen vor allem in der Eurokrise 2011 vor. Seither wird eine akkommodierende Geldpolitik nur noch zwei bis vier Mal erwähnt.

Inflationsdruck (inflationary pressures): Interessant: unter Trichet ein Thema, dann mit Draghi nicht mehr. Die Grafik ist sinnbildlich für die Schwerpunkte der beiden EZB-Chefs und könnte gut für einen Artikel zu Draghis Wirken verwendet werden, wenn er Herbst sein Präsidium abgibt.

Starke Wörter wie Krise oder Rezession werden vermieden, kommen sehr selten vor, ebenso Deflation.

Deflation kam nur einmal vor, im Intro vom Juli 2003 vor, noch unter dem Präsident Duisenberg vor. Das erstaunt, war doch die Furcht vor einem langfristigen Absigken des Preisniveuas (Deflation) der Hauptgrund für die unkonventionelle Geldpolitik (Anleihenkäufe).

Während das Stichwort Erholung (recovery) aus den Statements verschwunden ist, haben nun Wörter des Abschwungs Konjunktur (weak, slowdown)



Wie stark schaut die EZB auf die Finanzmärkte? Das ist eigentlich nicht Teil des Mandats. Doch in der Krise 2009 wurden die Finanzmärkte ein Dutzend Mal erwähnt. 2012 gab e seine Spitze von 6, seither nun noch ein bis zwei Mal.

Fazit: Die Resultate zum aktuellen Rand sind ähnlich wie in der Analyse der Monetary Accounts. Hinzu kommen interessante Ergebnisse über den längeren Zeitraum. Sie liefern Ideen für künftige EZB-Geschichten.

Textanalyse advanced/Maschine Learning

Similarity

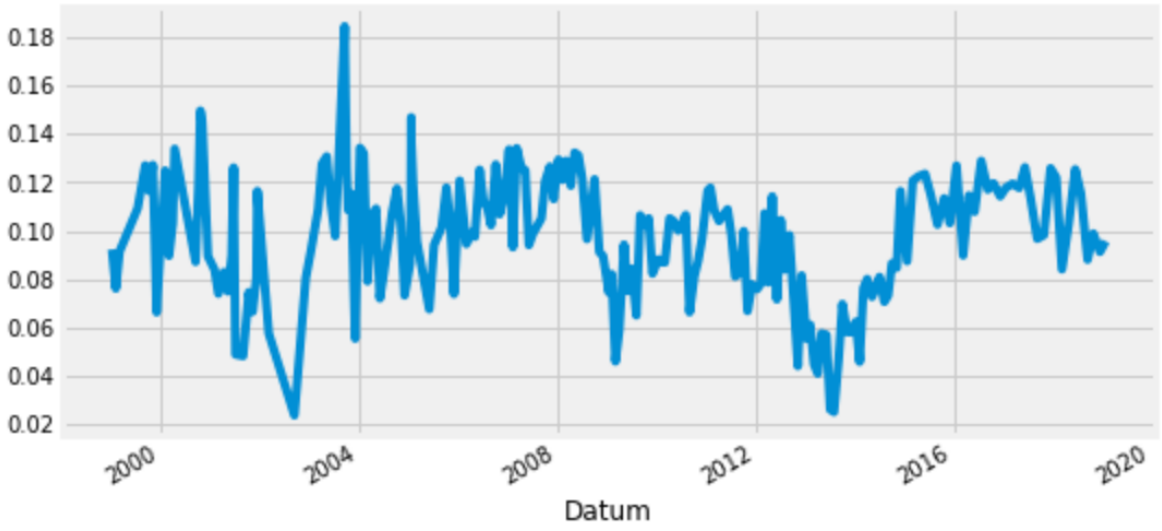
Auswertung mit Spacy:

Wie zu erwarten sind sich die Text sehr ähnlich, schliesslich werden sie sehr nach bestimmten Schema zusammengesetzt. Gegenüber dem Vormonat werden nur kleine Passagen geändert.

Die Ähnlichkeit zum jüngsten IS nimmt über die Jahre stetig zu. Auch das erstaunt nicht. Es gibt ansonsten keine Auffälligkeiten.

Sentiment

Auswertung mit TextBlob:



Auffällig ist, dass der allgemeine Ton in den Jahren 2013 und 2014 am negativsten war. In dieser Phase legte Draghi den Grundstein für die umstrittene, unkonventionelle Geldpolitik (Anleihenkäufe). Wie bei den Protokollen wurde der Ton jüngst wieder negativer.