ML alkalmazások projektfeladat

Péter Horváth

July 08, 2024

A projektfeladatban a Federal Reserve Bank által publikált Federal Open Market Committee (FOMC) ülések transcript-jeit elemzem NLP módszerekkel. Az elemzés célja, hogy feltárjam a kamatdöntő üléseken elhangzott témákat, és egy-egy indexet készítsek, amely időben leköveti ezen témák eloszlását minden egyes ülésre. A projekt a job market paper-öm első lépése, a kapott eredményeket monetáris politikai sokkok identifikációjára fogom használni, "továbbfejlesztve" Aruoba and Drechsel (2024) tanulmányát, amely hasonló ötleten alapszik, azonban a mögöttes szövegfeldolgozási módszerek egyszerűbbek.

A szükséges adatokat az FOMC oldalairól töltöttem le, ezek: https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm - amely az elmúlt 5 év üléseihez kapcsolódó dokumentumokat tartalmazza; valamint https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm - amely 1939-ig visszamenőleg tartalmazza ezen dokumentumokat (illetve azok elődjeit). A transcript-ek mellet számos más dokumentum típust is publikálnak, amelyek:

- Transcripts: Az ülés szó szerinti kézirata, amelyek mindig 5 éves csúszással kerülnek publikálásra.
- Memorandum of Discussions (Memo): A transcript-ekhez hasonló dokumentum, amely E/3-ban dokumentálja az ülésen elhangzottakat. A transcript-ekkel ellentétben azonban csak néhány évben kerültek publikálásra.
- Minutes: Ezek több néven is publikálásra kerültek az évek során: FOMC minutes (legfrissebb megnevezés), Historical Minutes és Records of Policy actions. A meeting minutes-ek egy rövidített összefoglalói az ülésen elhangzottaknak. A meeting minutes dokumentumok mindig az ülések után kerülnek publikálásra.
- Statements: Rövid (1-2 bekezdésnyi) összefoglalók a monetáris politikai döntésekről, melyek azonnal publikálásra kerülnek.
- "ColorBooks": A Red, Green, Blue, Teal és Beigebook kiadványokat tettem ebbe a kategóriába. Ezek gazdasági és pénzpiaci elemzések, amelyeket a FED Staff készít az ülések előtt (hasonlóak az MNB Inflációs jelenzéséhez).

Az alábbi grafikonon látható a kiadványtípusok időbeli alakulása:

Ebben az elemzésben csak a "Transcript" dokumentumokat használom, később a monetáris sokkok identifikációjához azonban hasznos lehet a "Colorbook"-ok használata is - pl.: a sokk identifikációjára egy

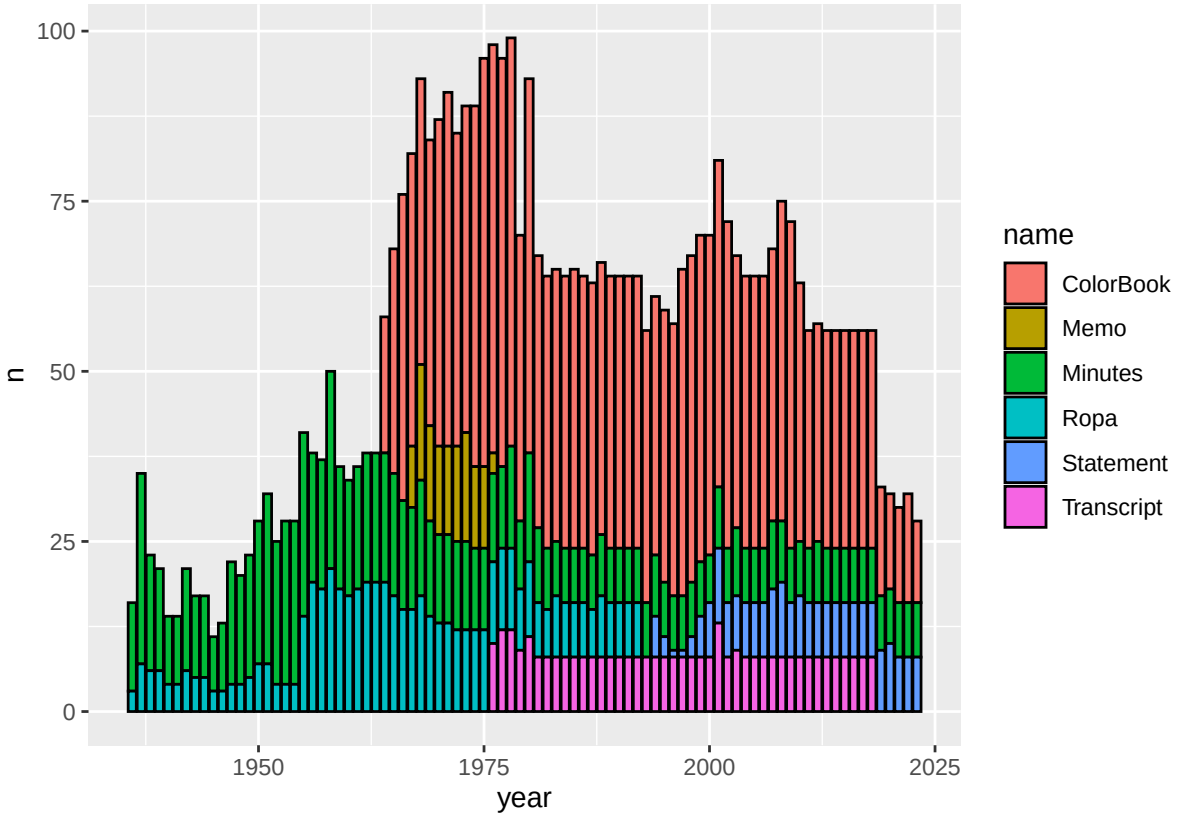

Figure 1: FOMC releases by publication type annually.

lehetséges megoldás a téma-eloszlások különbsége, vagy a témához kötött sentiment értékek különbsége az elemzők és a döntéshozók között.

A projektfeladatban végzett modellezésre Blei et al. (2003) LDA algoritmusát használom, Hansen et al. (2018) tanulmányához hasonlóan, akik szintén LDA modellt alkalmaznak az FOMC transcript-ek elemzéséhez. A feladathoz szükséges kódokat R-ben írtam, a következőkben pedig tételesen ismertetem a szövegfeldolgozás és LDA modell tanítás részleteit, végül pedig annak eredményeit.

Szövegfeldolgozás:

- A letöltött PDF file-okat bekezdésekre tagolom és regex-ek[1] segítségével eltávolítom belőle a fej- és lábléceket. Ennek eredményeképp minden nem-üres bekezdést véve hozok létre egy corpus-t.
- Második lépésként a dokumentumokban található szavakat part-of-speech tag-elem Wijffels (2021) UDPIPE implementációjának segítségével.
- A dokumentumokból egy-, két- és háromszavas (unigram, bigram, trigram) tokeneket képzek, amelyekből a POS tag segítségével specifikus szókapcsolatokat tartok csak meg. Unigram-ek esetén csak főneveket; Bigram-ek esetén melléknév-főnév és főnév-főnév kombinációk, Trigram-ek estén pedig melléknév-melléknév-főnév, melléknév-főnév-főnév, főnév-melléknév-főnév, főnév-főnév-főnév, valamint főnév-preopzíció-főnév. (Justeson and Katz (1995) tanulmánya ezeket azonosítja, mint specifikus szókapcsolat-fajták, amelyek képesek jól megragadni különböző témákat.)

[1]

- Végül a token-eket kisbetűsítem, eltávolítom a gyakran ismétlődő (stopword) szavakat, és minden szót annak lingvisztikai gyökerévé transzformálom (wordstem).

Token szelekció: Hansen et al. (2018) tanulmánya TF-IDF score alapján szűri a modell illesztéséhez használt token-eket, tapasztalatom szerint ez azonban i) túl szűkre csökkenti a token-ek számát, valamint, ii) eltávolít néhány gazdasági értelemben kulcsfontosságú token-t (pl. "inflat," "growth," "rate") azok alacsony IDF score-ja miatt. Egyszerűsítésképp, az elemzéshez három IDF cutoff értéket választok, ami azon token-eket szűri ki, amelyek a corpus dokumentumainak csak egy extrémen alacsony hányadában szerepelnek. A cutoff értékek 7.6 az unigram tokenek, 8.3 a bigram tokenek és 9 a trigram tokenek esetében. Az értékek rendre 0.05%, 0.025% and 0.0125% inverse document frequency hányadokra vonatkoznak.

LDA modell: Hansen et al. (2018) tanulmányától szintén eltérve, az unsupervised LDA algoritmus helyett egy semi-supervised seeded LDA algoritmust alkalmazok $k = 19$ témával, témánként 5 seedword-del, amelyek:

- Topic 1 - Inflation: 'inflat expect,' 'core inflat,' 'cpi,' 'price stabil,' 'inflationari pressur'
- Topic 2 - GDP growth: 'gdp growth,' 'real gdp,' 'output growth,' 'potenti gdp,' 'potenti output'
- Topic 3 - Labor Market: 'employ,' 'unemploy rate,' 'labor market,' 'real wage,' 'vacanc rate'
- Topic 4 - Monetary Policy: 'monetari polici,' 'polici decis,' 'quantit eas,' 'polici stanc,' 'forward guidanc'
- Topic 5 - Fiscal Policy: 'fiscal polici,' 'fiscal stimulu,' 'budget,' 'tax,' 'debt limit'
- Topic 6 - Banks: 'bank system,' 'loan,' 'credit,' 'financi institut,' 'leverag'
- Topic 7 - Yield Curve: 'yield curv,' 'term spread,' 'bond spread,' 'govern bond,' 'treasuri yield'
- Topic 8 - Housing Market: 'real estat,' 'mortgag rate,' 'properti,' 'hous price,' 'hous market'
- Topic 9 - Commodity Markets: 'oil,' 'food,' 'energi,' 'commod price,' 'natur ga'
- Topic 10 - Trade: 'trade,' 'import,' 'export,' 'current account,' 'good servic'
- Topic 11 - Exchange rates: 'currenc,' 'exchang,' 'appreci,' 'depreci,' 'foreign central bank'
- Topic 12 - Recessions: 'recess,' 'crisi,' 'downturn,' 'unemploy rise,' 'declin output'
- Topic 13 - Money Market: 'monei suppli,' 'monei growth,' 'monei market,' 'short rate,' 'credit growth'
- Topic 14 - Asset Purchase: 'balanc sheet,' 'asset purchas,' 'mb,' 'quantit eas,' 'open market oper'
- Topic 15 - Financial Stability: 'larg bank,' 'capit requir,' 'stress test,' 'system risk,' 'liquid risk'
- Topic 16 - Financial Markets: 'stock market,' 'equiti price,' 'corpor bond,' 'financi market volatil,' 'risk return'
- Topic 17 - Consumer Confidence: 'consum confid,' 'consum sentiment,' 'retail sale,' 'household spend,' 'auto sale'
- Topic 18 - Policy Rate: 'fund rate,' 'feder fund rate,' 'interest rate,' 'real rate,' 'natur rate'
- Topic 19 - Domestic Businesses: 'small busi,' 'inventori,' 'labor cost,' 'larg firm,' 'busi loan'

Hansen et al. (2018) $k = 40$ témával illesztett LDA modellt, így az optimális középpontszám megtalálása érdekében $k = [19, 40]$ intervallumon tanítok modelleket. Az végső k értéket pedgig perplexity és coherence

score-ok alapján választom. Az LDA modell hiperparamétereit $\alpha = 50/k$ és $\eta = 0.025$ kalibrálom a korábbi tanulmányt követve. Az LDA modelleket Gibbs sampling-gel becsülöm, 100 iterációval és 10 burnin-nel.

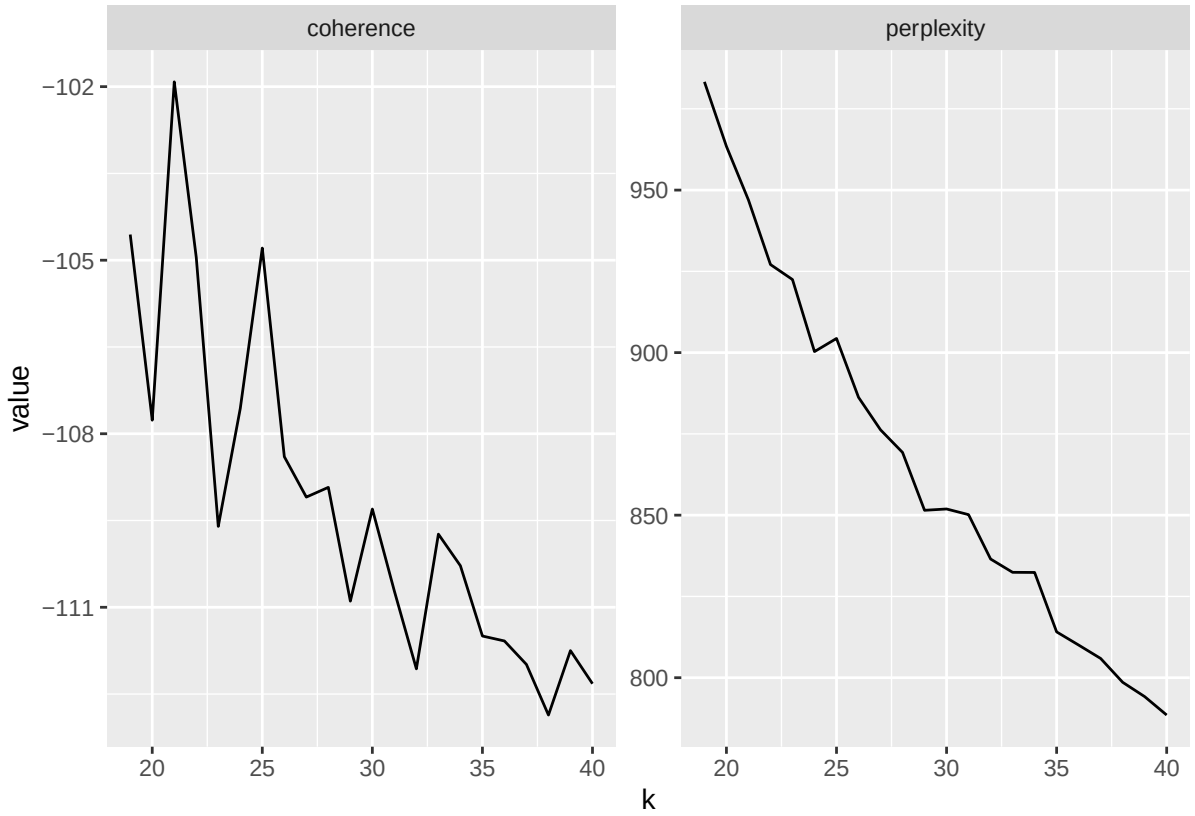

Figure 2: LDA evaluation metrics.

Nem meglepően, a perplexity score a lehető legnagyobb klaszter középpont számot javasolja, az átlagos topic-coherence score azonban a témák számával csökken, így $k = 20$[2] középponttal illesztem az elemzésre használt LDA modellt. A kapott modell β mátrixa megadja a token-ek témák közti eloszlását. Az alábbi grafikonon minden témához a 12 legvalószínűbb token-t ábrázolom.

[2]Kellő koherencia híjján a 20 témát 'Other'-ként címkéztem fel.

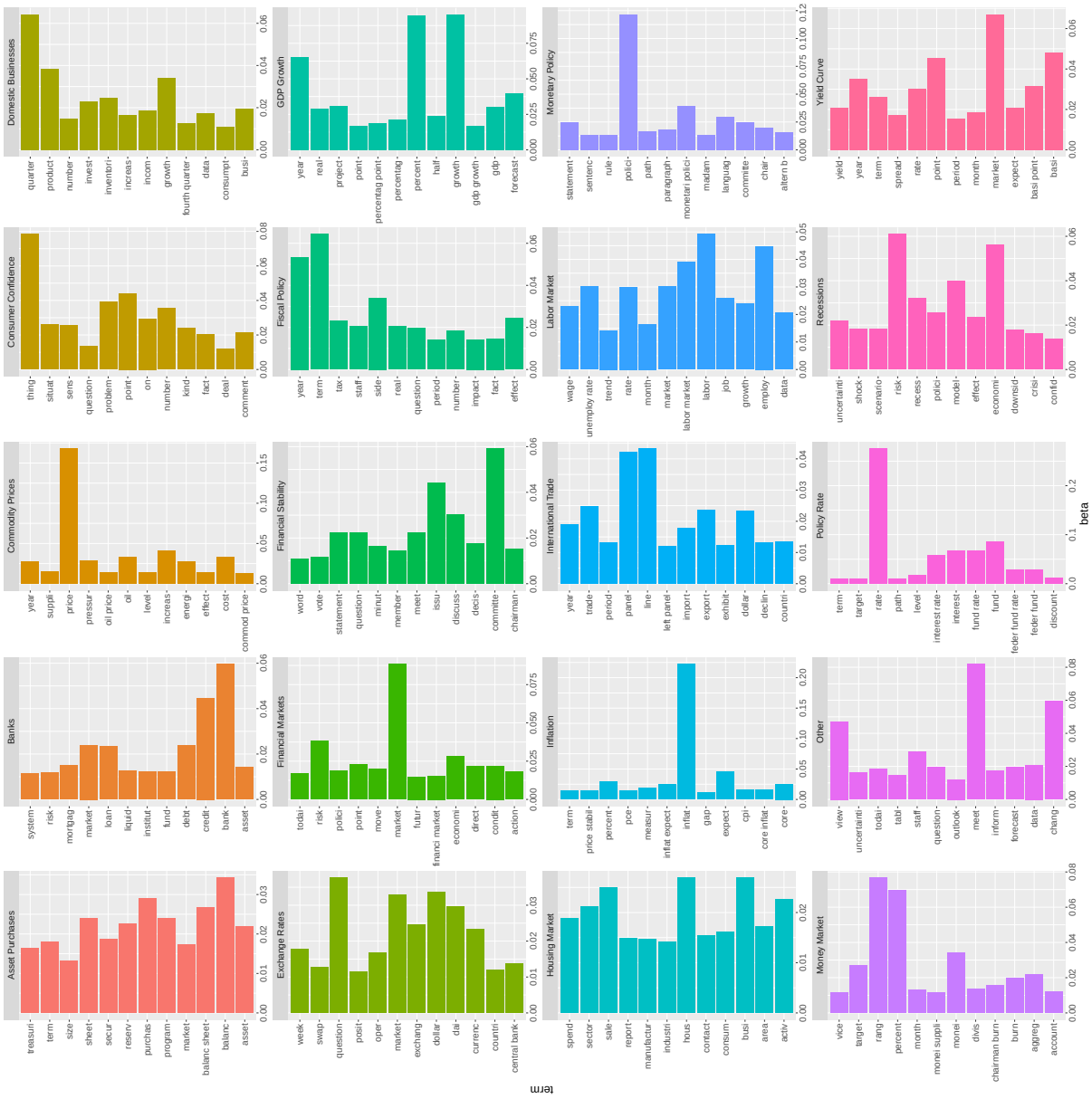

Figure 3: LDA evaluation metrics.

AZ LDA modellel továbbá megbecsülhető a dokumentumok téma-eloszlása is, melyet kiátlagolva származtatható, hogy egy adott publikációkban az egyes érintett témák az adott ülésben mennyire voltak hangsúlyosak. Az alábbi ábrán ennek időbeli alakulása látható.[3]

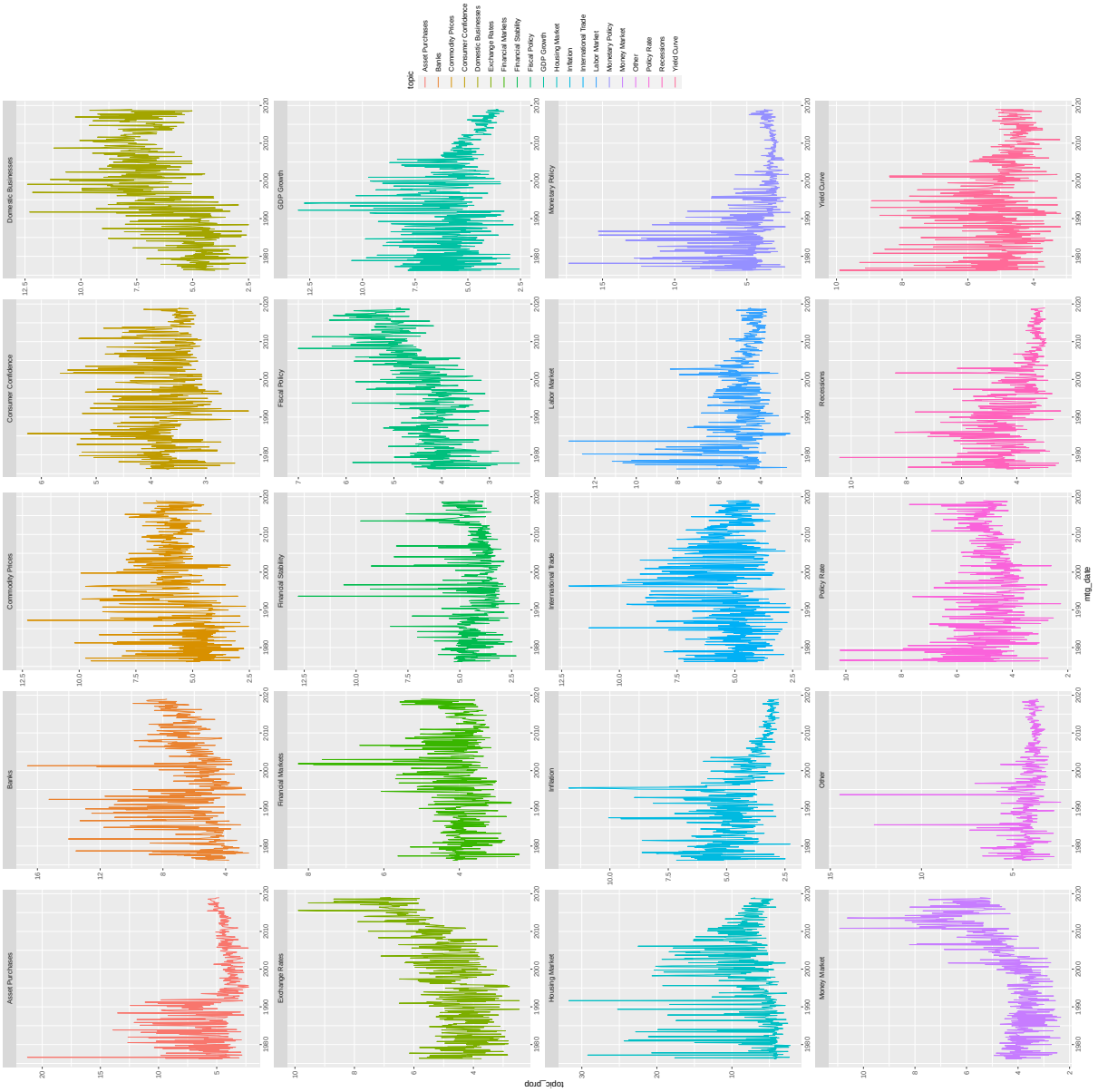

Figure 4: LDA evaluation metrics.

[3] A művelet ugyanúgy elvégezhető a többi publikációs típussal is - ezt elvégeztem, így az adatfile-jaimban ezek is megtalálhatók, itt csak a Transcript-ekkel készített idősort ábrázolom.

1 References

Aruoba, S Borağan, and Thomas Drechsel. 2024. "Identifying Monetary Policy Shocks: A Natural Language Approach." National Bureau of Economic Research.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133 (2): 801–70.

Justeson, John S, and Slava M Katz. 1995. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text." *Natural Language Engineering* 1 (1): 9–27.

Wijffels, Jan. 2021. "Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the'UDPipe' 'NLP'toolkit. 2020." *R Package Version 0.8* 8.