BRAIN

Brain Language Metrics on Company Filings – Version 2.1

Last Update 25/01/2023

File "stock_list_$date.csv"

The file contains the full list of stocks for which the 10-Ks and 10-Qs company reports are monitored daily. The file is marked with the export date *$date.*

Field	Type	Description	Example
COMPOSITE_FIGI	String	The FIGI composite code (https://www.openfigi.com) that uniquely identifies the stock across related exchanges in the same country	BBG000B9XRY4
TICKER	String	The stock ticker	AAPL
NAME	String	The company name	APPLE INC
SECTOR	String	The company sector	technology

File "metrics_*$reportCategory*_*$date*.csv"

The file contains the **language metrics** calculated on the company filings for the specified *$reportCategory*, for example 10-K. The *$reportCategory* "all" corresponds to the merge of 10-K (annual) and 10-Q (quarterly) reports. The file is marked with the export date *$date* and it is updated with daily frequency within 10 AM UTC.

Field	Type	Range	Description	Example
COMPOSITE_FIGI	String	-	The FIGI composite code (https://www.openfigi.com) that uniquely identifies the stock across related exchanges in the same country	BBG000B9XRY4
TICKER	String	-	The stock ticker	AAPL
DATE	String	-	The calculation date in YYYY-MM-DD format. The updated file for the current DATE is published every day within 12PM UTC.	2019-12-02
LAST_REPORT_CATEGORY	String	10-K or 10-Q	Category of last report (with respect to DATE) issued by the company.	10-K
LAST_REPORT_DATE	String	-	The date of last report (with respect to DATE) issued by the company in YYYY-MM-DD format	2019-08-31
N_SENTENCES	Number	[1, Inf]	Number of sentences extracted from the last available report.	1000
MEAN_SENTENCE_LENGTH	Number	[1, Inf]	The mean sentence length measured in terms of the mean number of words per sentence for the last available report.	50
SENTIMENT	Number	[-1,+1]	The financial sentiment of the last available report.	0.25

SCORE_UNCERTAINTY	Number	[0, 1]	The percentage of financial domain "*uncertainty"* language for present in the last report.	0.15
SCORE_LITIGIOUS	Number	[0, 1]	The percentage of financial domain *"litigious"* language for present in the last report.	0.15
SCORE_CONSTRAINING	Number	[0, 1]	The percentage of financial domain *"constraining"* language present in the last report.	0.1
SCORE_INTERESTING	Number	[0, 1]	The percentage of financial domain *"interesting"* language present in the last report.	0.1
READABILITY	Number	[0, Inf]	Reading grade level for the the report expressed by a number corresponding to US education grade. The score is obtained from the average of various readability tests to measure how difficult is the text to understand (e.g. Gunning Fog Index).	18.2
LEXICAL_RICHNESS	Number	[0, 1]	Lexical richness measured in terms of the Type-Token Ratio (TTR) which calculates the number of types (total number of words) divided by the number of tokens (number of unique words). The basic logic behind this measure is that if the text is more complex, the author uses a more varied vocabulary.	0.2
LEXICAL_DENSITY	Number	[0, 1]	Lexical density to measure the text complexity by computing the ratio between number of lexical words (nouns, adjectives, lexical verbs, adverbs) divided by the total number of words in the document.	0.5
SPECIFIC_DENSITY	Number	[0, 1]	Percentage of words belonging to the specific dictionary used for company filings analysis present in the last available report.	0.1
In the following we report the same metrics calculated above for the whole report for the **two sections "Risk Factors" (prefix RF) and "Management's Discussion and Analysis of Financial Condition and Results of Operations" (prefix MD)**. Both sections are present both in 10-K and 10-Q reports. In general if the fields are empty it means that the extraction of the specific section from the whole report failed or the result did not contain enough text to be able to calculate significant metrics.				
RF_N_SENTENCES	Number	[1, Inf]	Number of sentences extracted from the section "Risk Factors" of the last available report.	100
RF_MEAN_SENTENCE_LENGTH	Number	[1, Inf]	The mean sentence length measured in terms of the mean number of words per sentence for the section "Risk Factors" of the last available report.	50
RF_SENTIMENT	Number	[-1,+1]	The financial sentiment for the section "Risk Factors" of the last available report.	0.25
RF_SCORE_UNCERTAINTY	Number	[0, 1]	The percentage of financial domain "*uncertainty"* language present in the section "Risk Factors" of the last report.	0.15

RF_SCORE_LITIGIOUS	Number	[0, 1]	The percentage of financial domain *"litigious"* language present in the section "Risk Factors" of the last report.	0.15
RF_SCORE_CONSTRAINING	Number	[0, 1]	The percentage of financial domain *"constraining"* language present in the section "Risk Factors" of the last report.	0.1
RF_SCORE_INTERESTING	Number	[0,1]	The percentage of financial domain *"interesting"* language present in the section "Risk Factors" of the last report.	0.1
RF_READABILITY	Number	[0, Inf]	Reading grade level for the section "Risk Factors" of the last available report.	18.2
RF_LEXICAL_RICHNESS	Number	[0, 1]	Lexical richness for the section "Risk Factors" of the last available report.	0.2
RF_LEXICAL_DENSITY	Number	[0, 1]	Lexical density for the section "Risk Factors" of the last available report.	0.5
RF_SPECIFIC_DENSITY	Number	[0, 1]	Percentage of words belonging to the specific dictionary used for company filings analysis present in the section "Risk factors" of the last available report.	0.1
MD_N_SENTENCES	Number	[1, Inf]	Number of sentences extracted from the "MD&A" section of the last available report.	100
MD_MEAN_SENTENCE_LENGTH	Number	[1, Inf]	The mean sentence length measured in terms of the mean number of words per sentence for the "MD&A" section of the last available report.	50
MD_SENTIMENT	Number	[-1,+1]	The financial sentiment for the "MD&A" section of the last available report.	0.25
MD_SCORE_UNCERTAINTY	Number	[0, 1]	The percentage of financial domain "*uncertainty"* language present in the "MD&A" section of the last report.	0.15
MD_SCORE_LITIGIOUS	Number	[0, 1]	The percentage of financial domain *"litigious"* language present in the "MD&A" section of the last report.	0.15
MD_SCORE_CONSTRAINING	Number	[0, 1]	The percentage of financial domain *"constraining"* language present in the "MD&A" section of the last report.	0.1
MD_SCORE_INTERESTING	Number	[0, 1]	The percentage of financial domain *"interesting"* language present in the "MD&A" section of the last report.	0.1
MD_READABILITY	Number	[0, Inf]	Reading grade level for the "MD&A" section of the last available report.	18.2
MD_LEXICAL_RICHNESS	Number	[0, 1]	Lexical richness for the "MD&A" section of the last available report.	0.2
MD_LEXICAL_DENSITY	Number	[0, 1]	Lexical density for the "MD&A" section of the last available report.	0.5
MD_SPECIFIC_DENSITY	Number	[0, 1]	Percentage of words belonging to the specific dictionary used for company filings analysis present in the "MD&A" section of the last available report.	0.1

File "differences_*$reportCategory*_*$date*.csv"

The file contains the **differences of the language metrics** between the two last company reports for the specified *$reportCategory*, for example 10-K. The *$reportCategory* "all" corresponds to the merge of 10-K (annual) and 10-Q (quarterly) reports. The differences are always performed between reports of the same category and of the same period as explained below, for example:

- between annual 10-K of 2018 and annual 10-K of 2017;
- between quarterly 10-Q of first quarter 2019 and 10-Q of first quarter 2018.

The file is marked with the export date *$date* and it is updated with daily frequency within 10 AM UTC.

Field	Type	Range	Description	Example
COMPOSITE_FIGI	String	-	The FIGI composite code (https://www.openfigi.com) that uniquely identifies the stock across related exchanges in the same country.	BBG000B9XRY4
TICKER	String	-	The stock ticker	AAPL
DATE	String	-	The calculation date in YYYY-MM-DD format. The updated file for the current DATE is published every day within 12PM UTC.	2019-12-02
LAST_REPORT_DATE	String	-	The date of last report (with respect to DATE) issued by the company in YYYY-MM-DD format.	2019-08-31
LAST_REPORT_CATEGORY	String	10-K or 10-Q	Category of last report (with respect to DATE) issued by the company.	10-K
LAST_REPORT_PERIOD	Number	10-K: [1, ..N] 10-Q: [1,3]	The period of the last available report. For 10-K annual reports this is an integer number labelling the annual reports. For 10-Q quarterly reports this an integer number from 1 to 3 labelling the period report. This is used to perform differences between reports of the same period.	2
PREV_REPORT_DATE	String	-	The date of previous report (with respect to LAST_REPORT_DATE) issued by the company in YYYY-MM-DD format.	2018-08-31
PREV_REPORT_CATEGORY	String	10-K or 10-Q	Category of previous report (with respect to LAST_REPORT_DATE) issued by the company.	10-K
PREV_REPORT_PERIOD	Number	10-K: [1, ..N] 10-Q: [1,3]	The period of the last available report. For 10-K annual reports this is an integer number labelling the annual reports. For 10-Q quarterly report this a integer number from 1 to 3 labelling the period report. This is used to perform differences between reports of the same period.	3
DELTA_PERC_N_SENTENCES	Number	[-Inf,Inf]	Percentage change of the number of sentences between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1

DELTA_PERC_MEAN_SENTENCE_LENGTH	Number	[-Inf,Inf]	Percentage change of sentence length (mean number of words per sentence) between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.2
DELTA_SENTIMENT	Number	[-2,+2]	The difference of financial sentiment between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1
DELTA_SCORE_UNCERTAINTY	Number	[-1,+1]	The difference of percentage of financial domain "*uncertainty"* language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1
DELTA_SCORE_LITIGIOUS	Number	[-1,+1]	The difference of percentage of financial domain "*litigious"* language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1
DELTA_SCORE_CONSTRAINING	Number	[-1,+1]	The difference of percentage of financial domain "*constraining"* language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1
DELTA_SCORE_INTERESTING	Number	[-1,+1]	The difference of percentage of financial domain "*interesting"* language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.1
DELTA_READABILITY	Number	[-Inf, Inf]	The difference of the readability metric between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	-0.5
DELTA_LEXICAL_RICHNESS	Number	[-1,+1]	The difference of the lexical richness metric between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.05
DELTA_LEXICAL_DENSITY	Number	[-1,+1]	The difference of the lexical density metric between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.05
DELTA_SPECIFIC_DENSITY	Number	[-1,+1]	The difference of the specific density metric between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.05
SIMILARITY_ALL	Number	[0, 1]	The language similarity between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8

SIMILARITY_POSITIVE	Number	[0, 1]	The similarity in terms of financial domain "*positive*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
SIMILARITY_NEGATIVE	Number	[0, 1]	The similarity in terms of financial domain "*negative*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
SIMILARITY_UNCERTAINTY	Number	[0, 1]	The similarity in terms of financial domain "*uncertainty*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
SIMILARITY_LITIGIOUS	Number	[0, 1]	The similarity in terms of financial domain "*litigious*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
SIMILARITY_CONSTRAINING	Number	[0, 1]	The similarity in terms of financial domain "*constraining*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
SIMILARITY_INTERESTING	Number	[0, 1]	The similarity in terms of financial domain "*interesting*" language between the last available report (LAST_REPORT_DATE) and the previous report of same period and category (PREV_REPORT_DATE).	0.8
In the following we report the same metrics calculated above for the whole report for the **two sections "Risk Factors" (prefix RF) and "Management's Discussion and Analysis of Financial Condition and Results of Operations" (prefix MD)**. Both sections are present both in 10-K and 10-Q reports. In general if the fields are empty it means that the extraction of the specific section failed for one of the two compared reports or the result did not contain enough text to be able to calculate significant metrics for one of the two compared reports. For the specific sections we only report the similarity for "all", "positive" and "negative" language.				
RF_DELTA_PERC_N_SENTENCES	Number	[-Inf, Inf]	Percentage change of the number of sentences between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
RF_DELTA_PERC_MEAN_SENTENCE_LENGTH	Number	[-Inf, Inf]	Percentage change of mean sentence length between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.2
RF_DELTA_SENTIMENT	Number	[-2,+2]	The difference of financial sentiment between the "Risk Factors" sections of the last available report	-0.1

			(LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	
RF_DELTA_SCORE_UNCERTAINTY	Number	[-1,+1]	The difference of percentage of financial domain "*uncertainty*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
RF_DELTA_SCORE_LITIGIOUS	Number	[-1,+1]	The difference of percentage of financial domain "*litigious*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
RF_DELTA_SCORE_CONSTRAINING	Number	[-1,+1]	The difference of percentage of financial domain "*constraining*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
RF_DELTA_SCORE_INTERESTING	Number	[-1,+1]	The difference of percentage of financial domain "*interesting*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
RF_DELTA_READABILITY	Number	[-Inf,Inf]	The difference of the readability metric between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.5
RF_DELTA_LEXICAL_RICHNESS	Number	[-1,+1]	The difference of the lexical richness metric between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.05
RF_DELTA_LEXICAL_DENSITY	Number	[-1,+1]	The difference of the lexical density metric between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.05
RF_DELTA_SPECIFIC_DENSITY	Number	[-1,+1]	The difference of the specific density metric between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same	0.05

			section of the previous report of same period and category (PREV_REPORT_DATE).	
RF_SIMILARITY_ALL	Number	[0, 1]	The language similarity between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.8
RF_SIMILARITY_POSITIVE	Number	[0, 1]	The similarity in terms of financial domain "*positive*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.8
RF_SIMILARITY_NEGATIVE	Number	[0, 1]	The similarity in terms of financial domain "*negative*" language between the "Risk Factors" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.8
MD_DELTA_PERC_N_SENTENCES	Number	[-Inf,Inf]	Percentage change of the number of sentences between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
MD_DELTA_PERC_MEAN_SENTENCE_LENGTH	Number	[-Inf,Inf]	Percentage change of mean sentence length between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.2
MD_DELTA_SENTIMENT	Number	[-2,+2]	The difference of financial sentiment between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
MD_DELTA_SCORE_UNCERTAINTY	Number	[-1,+1]	The difference of percentage of financial domain "*uncertainty*" language between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
MD_DELTA_SCORE_LITIGIOUS	Number	[-1,+1]	The difference of percentage of financial domain "*litigious*" language between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of	-0.1

			same period and category (PREV_REPORT_DATE).	
MD_DELTA_SCORE_CONSTRAINING	Number	[-1,+1]	The difference of percentage of financial domain "*constraining"* language between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
MD_DELTA_SCORE_INTERESTING	Number	[-1,+1]	The difference of percentage of financial domain "*interesting"* language between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.1
MD_DELTA_READABILITY	Number	[-Inf, Inf]	The difference of the readability metric between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	-0.5
MD_DELTA_LEXICAL_RICHNESS	Number	[-1,+1]	The difference of the lexical richness metric between the "MD&A" sections of the last available report and the same section of the previous report of the same period and category.	0.05
MD_DELTA_LEXICAL_DENSITY	Number	[-1,+1]	The difference of the lexical density metric between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.05
MD_DELTA_SPECIFIC_DENSITY	Number	[-1,+1]	The difference of the specific density metric between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.05
MD_SIMILARITY_ALL	Number	[0, 1]	The language similarity between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.8
MD_SIMILARITY_POSITIVE	Number	[0, 1]	The similarity in terms of financial domain "*positive"* language between the "MD&A" sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	0.8
MD_SIMILARITY_NEGATIVE	Number	[0, 1]	The similarity in terms of financial domain "*negative"* language	0.8

			between the “MD&A” sections of the last available report (LAST_REPORT_DATE) and the same section of the previous report of same period and category (PREV_REPORT_DATE).	

Contacts

For more information please contact *support@braincompany.co*

Disclaimer

The content of this document is not to be intended as investment advice. The material is provided for informational purposes only and does not constitute an offer to sell, a solicitation to buy, or a recommendation or endorsement for any security or strategy, nor does it constitute an offer to provide investment advisory or other services by Brain. Brain makes no guarantees regarding the accuracy and completeness of the information expressed in this document