



```

1 . cd "C:\Users\eliven\Dropbox\ELLW_2026\code"
  C:\Users\eliven\Dropbox\ELLW_2026\code

2 . doedit "03_other_dataset_construction.do"

3 . do "C:\Users\eliven\Dropbox\ELLW_2026\code\03_other_dataset_construction.do"

4 . *****
> *****
5 . *****
> *****
6 . *****
> *****
7 .
8 . * Construct firm-year dataset, industry dataset, and guidance (merge version) dataset
9 .
10 . *****
> *****
11 .
12 . * Set directory
13 . global dataset "datasets"

14 .
15 . *****
> *****
16 . * Firm-year dataset
17 .
18 . * Import firm-date dataset
19 . use "$dataset\firm_date", replace

20 .
21 . * Aggregate daily query and task metrics into firm-level annual totals
22 . collapse (sum) info_query_q info_analysis_q shareholder_change_q product_tech_innov_q /
> //
> customer_brand_q macro_env_q supply_chain_q regulation_policy_q restructuring_debt_q //
> /
> invest_decision_q finance_decision_q production_ops_q exec_team_q internal_control_q //
> /
> talent_mgmt_culture_q csr_q market_comp_q business_outlook_q financial_perf_q ///
> stock_perf_q listing_delisting_q bank_q renewable_energy_q ai_q pharma_q ///
> media_entertain_q education_q specific_q gen_task_q edu_task_q info_proc_q info_sense_q
> ///
> stock_rec_q event_mon_q other_sense_q info_search_q cur_info_q hist_info_q ///
> search_qual_q search_quant_q src_report_q src_announce_q src_investor_q ///
> src_othercorp_q src_mediary_q src_regulator_q src_otherplat_q info_verify_q ///
> verif_report_q verif_announce_q verif_investor_q verif_othercorp_q verif_mediary_q ///
> verif_regulator_q verif_otherplat_q cause_anal_q summary_q compare_q sameco_time_q ///
> trend_pred_q trend_qual_q trend_quant_q co_trend_q ind_trend_q impact_anal_q ///
> co_impact_q ind_impact_q macro_impact_q overall_eval_q fs_anal_q risk_anal_q general_co
> unt industry_count stock_rec_c ind_stockrec_co ///
> strat_anal_q other_anal_q, by(stkcd)

23 .
24 . * Export firm-year dataset
25 . save "$dataset\firm_year", replace
    file datasets\firm_year.dta saved

26 .
27 . *****
> *****

```

```

28 . * Industry dataset
29 .
30 . * Load original dataset and keep controls of 2023
31 . use "$dataset\merged_yearly_controls.dta", clear

32 . keep if year == 2023
    (17,663 observations deleted)

33 .
34 . * Convert string industry codes into numeric labels
35 . encode industrycodec, gen(industry)

36 . keep stkcd industry

37 . save "$dataset\industry", replace
    file datasets\industry.dta saved

38 .
39 . *****
    > *****
40 . * Guidance (merge version) dataset
41 .
42 . * 1. Data Cleaning and Formatting
43 . import delimited "$dataset\mgt_guidance.csv",clear
    (encoding automatically selected: UTF-8)
    Note: Unmatched quote while processing row 541; this can be due to a formatting problem
          in the file or because a quoted data element spans multiple lines. You should
          carefully inspect your data after importing. Consider using option
          bindquote(strict) if quoted data spans multiple lines or option bindquote(nobind)
          if quotes are not used for binding data.
    Note: Unmatched quote while processing row 541; this can be due to a formatting problem
          in the file or because a quoted data element spans multiple lines. You should
          carefully inspect your data after importing. Consider using option
          bindquote(strict) if quoted data spans multiple lines or option bindquote(nobind)
          if quotes are not used for binding data.
    Note: Unmatched quote while processing row 549; this can be due to a formatting problem
          in the file or because a quoted data element spans multiple lines. You should
          carefully inspect your data after importing. Consider using option
          bindquote(strict) if quoted data spans multiple lines or option bindquote(nobind)
          if quotes are not used for binding data.
    (228 vars, 2,093 obs)

44 . drop v1

45 . drop _merge

46 . destring scode, replace force
    scode: contains nonnumeric characters; replaced as long
    (8 missing values generated)

47 . rename scode stkcd

48 .
49 . * Convert string publication date to Stata date format
50 . gen date1 = date(publdate,"YMD")
    (6 missing values generated)

51 . drop date

52 . rename date1 date_guidance

```

```

53 .
54 . * Remove duplicate entries for the same firm-date to ensure a clean merge
55 . bys date_guidance stkcd: gen index = _N

56 . drop if index > 1
    (12 observations deleted)

57 . drop index

58 . save "$dataset\guidance", replace
    file datasets\guidance.dta saved

59 .
60 . * 2. Feature Engineering: Content and Reason Topic Intensity
61 . use "$dataset\guidance", replace

62 . drop pure_content

63 .
64 . * Step 1: Identify all variables that end with _content and _reason
65 . ds *_content
    股东回报与~t      公司投资决~t      市场竞争_c~t      ai_content
    产品技术与~t      公司融资决~t      公司商业前~t      医药_content
    顾客与品牌~t      公司生产与~t      公司财务表~t      传媒与娱乐~t
    宏观经济环~t      董监高管理~t      公司股市表~t      教育行业_c~t
    供应链_con~t      公司内部控~t      上市与退市~t
    监管与政策~t      人才管理与~t      银行_content
    企业重整与~t      csr_content      新能源_con~t

66 . local content_vars = r(varlist)

67 .
68 . ds *_reason
    股东回报与~n      公司投资决~n      市场竞争_r~n      ai_reason
    产品技术与~n      公司融资决~n      公司商业前~n      医药_reason
    顾客与品牌~n      公司生产与~n      公司财务表~n      传媒与娱乐~n
    宏观经济环~n      董监高管理~n      公司股市表~n      教育行业_r~n
    供应链_rea~n      公司内部控~n      上市与退市~n
    监管与政策~n      人才管理与~n      银行_reason
    企业重整与~n      csr_reason      新能源_rea~n

69 . local reason_vars = r(varlist)

70 .
71 . * Step 2: Create a new variable content_topics that counts how many _content variables
    > are greater than 0 for each observation
72 . gen content_topics = 0

73 . foreach var of local content_vars {
    2.     replace content_topics = content_topics + (`var' > 0)
    3. }
    (1,308 real changes made)
    (4 real changes made)
    (5 real changes made)
    (4 real changes made)
    (4 real changes made)
    (17 real changes made)
    (835 real changes made)
    (9 real changes made)
    (3 real changes made)
    (22 real changes made)
    (2 real changes made)
    (32 real changes made)
    (1 real change made)
    (0 real changes made)
    (4 real changes made)
    (373 real changes made)
    (2,081 real changes made)
    (130 real changes made)
    (13 real changes made)
    (2 real changes made)
    (9 real changes made)

```

```
(2 real changes made)
(62 real changes made)
(0 real changes made)
(2 real changes made)
```

```
74 .
75 . * Step 3: Create a new variable reason_topics that counts how many _reason variables ar
    > e greater than 0 for each observation
76 . gen reason_topics = 0

77 . foreach var of local reason_vars {
    2.     replace reason_topics = reason_topics + (`var' > 0)
    3. }
(583 real changes made)
(1,415 real changes made)
(708 real changes made)
(951 real changes made)
(451 real changes made)
(455 real changes made)
(383 real changes made)
(770 real changes made)
(171 real changes made)
(1,983 real changes made)
(156 real changes made)
(728 real changes made)
(265 real changes made)
(19 real changes made)
(1,107 real changes made)
(1,283 real changes made)
(2,033 real changes made)
(41 real changes made)
(23 real changes made)
(106 real changes made)
(314 real changes made)
(152 real changes made)
(194 real changes made)
(9 real changes made)
(20 real changes made)

78 .
79 . * Step 4: Optionally, label the new variables
80 . label variable content_topics "Count of _content variables > 0"

81 . label variable reason_topics "Count of _reason variables > 0"

82 .
83 . * 3. Final variable construction and selection
84 . gen financial_reason = 公司财务表现_reason_word

85 . gen production_reason = 公司生产与运营_reason_word

86 . gen tech_reason = 产品技术与创新_reason_word

87 . gen competition_reason = 市场竞争_reason_word

88 . gen business_outlook_reason = 公司商业前景_reason_word

89 .
90 . rename date_guidance date
```

```
91 .
92 . * Create a dummy indicator for the presence of Management Forecast
93 . gen mef_text = 1

94 .
95 . * Retain only the necessary variables for the final merging process
96 . keep *_topics total_sentence reason_sentence_ratio total_word date stkcd reason_word fi
    > nancial_reason tech_reason production_reason competition_reason business_outlook_reaso
    > n mef_text

97 .
98 . * Export the dataset to merge
99 . save "$dataset\guidance_formerge", replace
    file datasets\guidance_formerge.dta saved

100 .
    end of do-file

101 .
```