



银行业景气度与数字化转型分析

队名：津京乐道队

队员：肖俊成（天津大学，金融专硕研二）

于仰志（天津大学，金融专硕研二）

廖嘉慧（天津大学，金融专硕研二）

刘菁泽（天津大学，金融专硕研二）

苏聪（天津大学，金融专硕研二）

报告日期：2023/4/25

目录

1. 银行业概览	1
1.1 银行业简介	1
1.2 银行龙头情况	2
1.3 未来发展趋势：数字化转型，金融科技助银行业高质量发展	2
1.3.1 国家政策帮扶银行业数字化转型	2
1.3.2 央行以及各银行为数字化转型与科技助力高质量发展所做的努力	2
2. 银行竞争格局	3
2.1 “一股独大”不复存在	3
2.2 区域差异显著	4
2.3 数字化转型的破局效应	5
3. 数字化转型的量化分析	5
3.1 数字化转型的应用领域	5
3.2 数字化转型的指标构造——基于文本分析	6
4. 银行业景气度的定量刻画	7
4.1 供需景气度指标筛选	7
4.1.1 表观供需差异	7
4.1.2 潜在需求增速	8
4.2 银行业供需景气度指标构造	8
4.3 景气度表现	9
5. 数字化转型与行业表现	10
6. 数字化转型下的行业展望	12
7. 数字化转型风险提示	14
附录	14
A. 参考文献	14
B. 金融科技子公司成立情况	15
C. 银行数字化相关战略	15
D. 网络爬虫及文本分析相关代码	16

图表目录

图 1 : 银行业上市公司分布	1
图 2 : 2021-2022 前十银行净利润 (单位: 亿元)	2
图 3 : 2021-2022 前十银行营业收入 (单位: 亿元)	2
图 4 : 金融科技投入情况	3
图 5 : 2019-2022 年各类银行资产规模变化 (单位: 万亿元)	3
图 6 : 各省份的银行竞争水平	4
图 7 : 银行数字化转型供需端	5
图 8 : 文本分析中的关键词说明	6
图 9 : 底层技术关键词词频 (平均每家企业)	7
图 10 : 应用领域关键词词频 (平均每家企业)	7
图 11 : 金融机构各项贷款余额与存款余额	7
图 12 : 金融机构贷存差与行业指数	7
图 13 : 房地产开发投资完成额累计值 (yoy) 与贷存差 (yoy) 相关性分析	8
图 14 : 固定资产投资完成额累计值 (yoy) 与贷存差 (yoy) 相关性分析	8
图 15 : 金融机构贷存差 yoy 与行业指数相关性分析	8
图 16 : 供需景气度-备选与行业指数相关性分析	8
图 17 : 行业景气度与申万银行指数	9
图 18 : 行业景气度与数字化技术总词频	10
图 19 : 数字化技术总词频与银行行业营业收入增速	10
图 20 : 数字化技术总词频与银行行业净利润增速	10
图 21 : 数字化技术总词频与不良贷款率	11
图 22 : 数字化技术总词频与核心一级资本充足率	11
图 23 : 行业景气度与数字化转型五维度词频	12
表 1 : 数字化转型的技术与应用	5
表 2 : 数字化转型各维度词频	10
表 3 : 数字化技术各维度与行业景气度相关性	12
表 4 : 2021-2022 年人工智能技术词频与大数据技术词频 TOP5	12
表 5 : 邮储银行、平安银行、上海银行指标统计 (单位: 亿元、%)	13
表 6 : 金融科技子公司成立情况	15
表 7 : 银行数字化相关战略	15

1. 银行业概览

1.1 银行业简介

我国银行体系主要由一个中央银行、三个政策性银行、国有商业银行、股份制商业银行、城商行、农信社、以及民营银行以及外资银行组成。

(1) 中央银行：中国人民银行。央行不对个人或企业办理业务，主要职能为执行中央货币政策，对国民经济进行宏观调控，对金融机构乃至金融业进行监督管理；

(2) 政策性银行：包括国家开发银行、中国进出口银行、中国农业发展银行。其不以营利为目的，专门为贯彻、配合政府社会经济政策或意图，在特定的业务领域内，直接或间接地从事政策性融资活动，充当政府发展经济、促进社会进步、进行宏观经济管理的工具；

(3) 国有商业银行：包括工商银行、农业银行、中国银行、建设银行、交通银行等。由财政部、中央汇金公司直接管控，并且具有盈利性质。其主要职责为通过存款、贷款、汇兑、储蓄等业务，承担信用中介的金融机构；

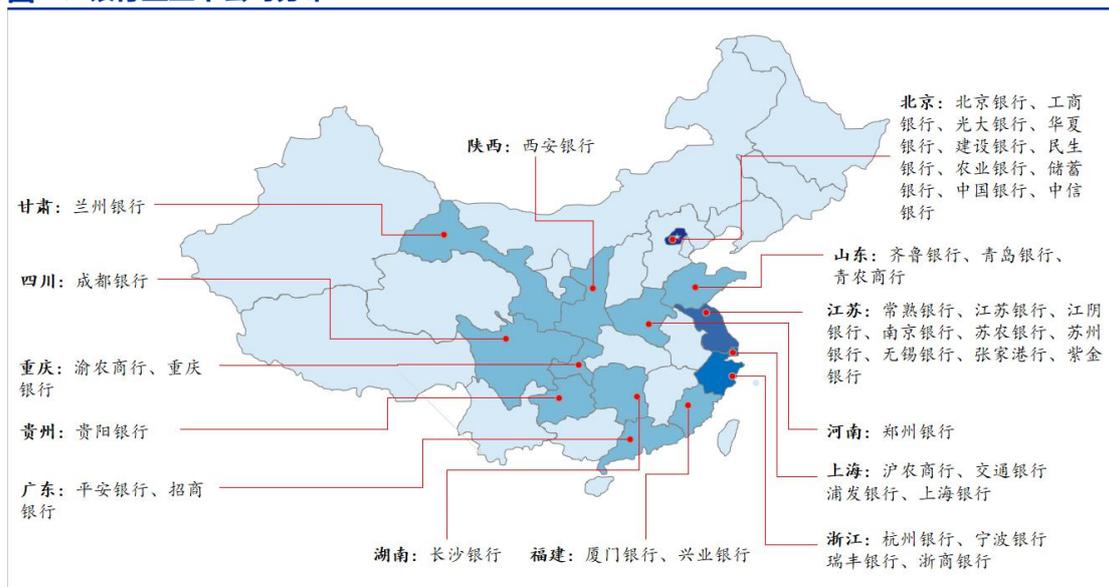
(4) 股份制商业银行：包括招商银行、浦发银行、中信银行、民生银行等。其和国有商业银行的区别在于，股份制商业银行中的国有资本占比较小；

(5) 城商行：包括北京银行、上海银行、江苏银行等，具有明显的区域性特征，主要为中小企业提供金融服务；

(6) 农村信用社：主要职能为筹集闲散资金，为农村和农村的经济发展提供金融服务；

(7) 民营银行与外资银行。其中民营银行为个人发起的银行，包括前海微众银行、民商银行、金城银行等，外资银行包括汇丰银行、花旗银行、渣打银行、东亚银行等。

图 1：银行业上市公司分布



资料来源：小组成员整理

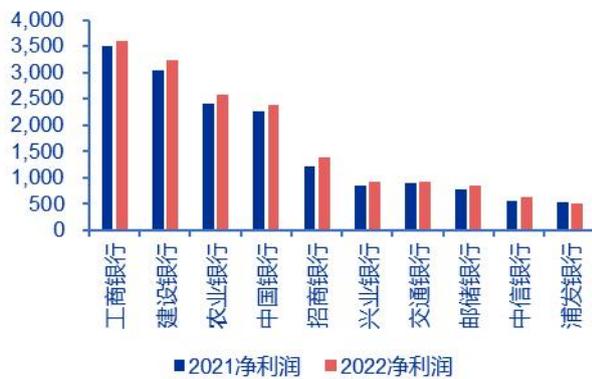
从银行业上市公司注册地来看，我国银行业目前共有 42 家 A 股上市公司，其中东部地

区的银行业上市公司较多，其中北京市 10 家、江苏省 9 家。北京市银行业上市公司主要有工商银行、建设银行、中国银行、农业银行等，江苏省银行业上市公司主要有南京银行、江苏银行、无锡银行、苏农银行等。此外，我国股份制商业银行中规模靠前的两家银行均位于广东省，分别为招商银行、平安银行。我国其余地区的银行业上市公司主要以地方银行为主。

1.2 银行龙头情况

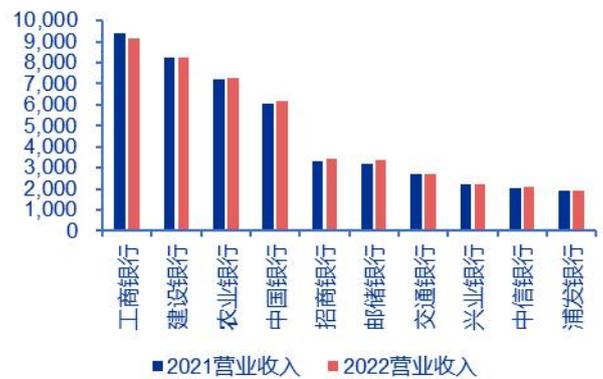
从近两年业绩来看，2022 年我国银行业净利润前十名的银行分别是工商银行、建设银行、农业银行、中国银行、招商银行、兴业银行、交通银行、邮储银行、中信银行、浦发银行，与 2021 年相比，仅兴业银行和交通银行浮动 1 个位次。同时 2022 年营业收入排名前十依然是上述银行，且与 2021 年排名一致。总体来说，近年银行业龙头格局较为稳定。

图 2: 2021-2022 前十银行净利润 (单位: 亿元)



资料来源: Wind

图 3: 2021-2022 前十银行营业收入 (单位: 亿元)



资料来源: Wind

1.3 未来发展趋势: 数字化转型, 金融科技助银行业高质量发展

1.3.1 国家政策帮扶银行业数字化转型

2021 年 12 月中国人民银行发布的《金融科技发展规划(2022—2025 年)》和 2022 年银保监会印发的《关于加强商业银行互联网贷款业务管理 提升金融服务质效的通知》和《关于银行业保险业数字化转型的指导意见》，上述文件要求银行加强金融科技治理能力，不断完善现代化科技治理架构。金融科技是技术驱动的金融创新，我国银行业作为经济生活中的重要成员，需要稳妥发展金融科技，加快金融机构数字化转型，进一步推动金融科技迈入高质量发展的新阶段，更充分发挥金融科技赋能作用，增强金融服务实体经济的能力和效率。

1.3.2 央行以及各银行为数字化转型与科技助力高质量发展所做的努力

央行为完善数据共享、数字货币清结算、征信数据库建设等工作，正稳步推进计划中的金融科技工作。目前兴业银行等 21 家大型银行都已成立自己的金融科技子公司，具体情况见附录 B。除了以上银行的金融科技子公司外，许多银行自身也采取了一些数字化战略，对

自身的组织架构进行了升级更新，具体情况见附录 C。

同时，我国商业银行也在不断地加大金融科技方面的资金投入力度与发展速度。下图选取了几家具有较高金融科技资金投入的代表性银行，可以看到这些银行的金融科技投入资金较多，且仍保持在一个较高的增长水平，这说明我国银行业普遍重视数字化转型，且认识到了金融科技能力在数字化转型过程中的决定性作用。

图 4：金融科技投入情况



资料来源：公开资料整理

2. 银行竞争格局

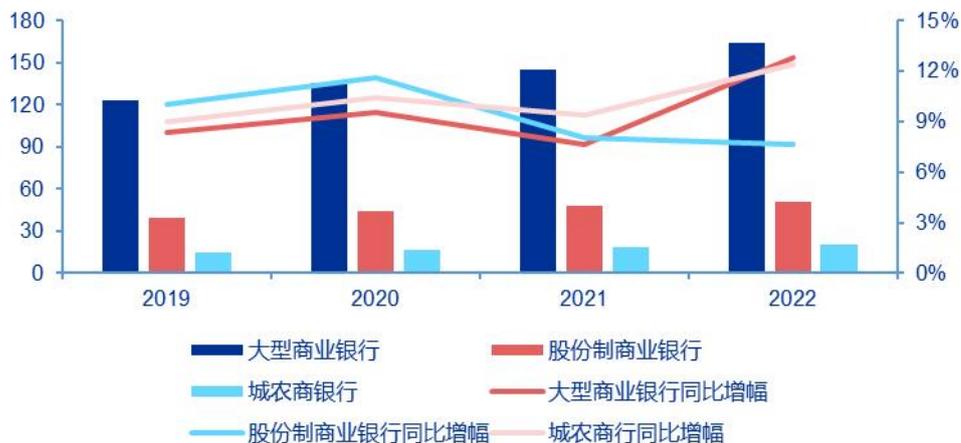
2.1 “一股独大”不复存在

自 2001 年中国加入 WTO 后，外资银行得以进入中国市场，打破原有的国有银行垄断格局，银行竞争程度逐年加大，银行业市场化进程也不断加快。2006 年，银监会放开了农村地区银行业市场准入，并允许城市地区商业银行可依据自身发展情况设立异地分支机构，各地银行机构数量显著增长，银行竞争程度明显加大。

2009 年 4 月，中国银监会再次加大对于股份制商业银行和城市商业银行的异地准入力度，放松对股份制和城市商业银行在异地设立分支机构的限制，设立分支机构数量将不再受到数量控制。大量股份制商业银行和城市商业银行分支机构被批准设立，各类非国有商业银行分支机构数量也在逐年增加，银行竞争程度进一步加大。

2015 年，银行业全面打开对民间资本的大门，符合条件的民营企业被允许以自有资金投资设立银行业金融机构。同时，商业银行和农村合作金融机构存款利率浮动上限也于 2015 年正式取消，银行市场差异化竞争加剧。

图 5：2019-2022 年各类银行资产规模变化 (单位：万亿元)



资料来源：CSMAR 数据库

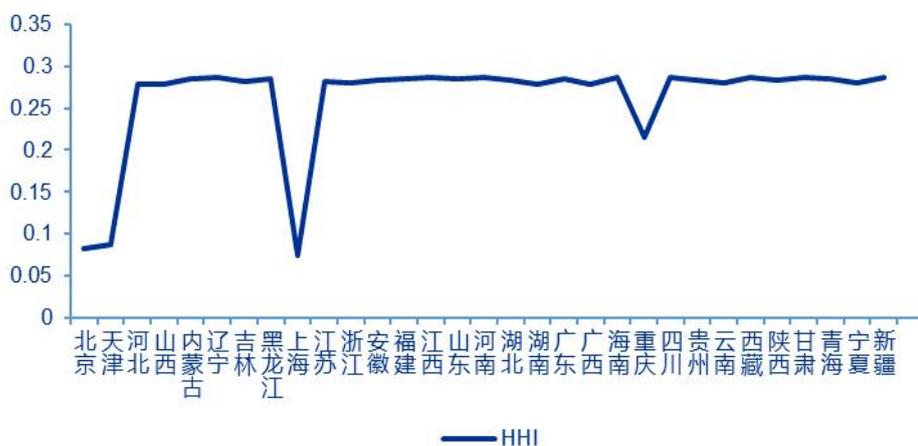
大型商业银行的同比增幅在 2019-2021 年间不超过 10%，低于城农商行和股份制银行。截止 2022 年 12 月，6 家国有商业银行资产市场总额超 164.11 万亿元，份额为 65.12%，而股份制商业银行和农商行的资产市场份额之和达到 34.88%。国有大型商业银行长期以来“一股独大”的垄断市场地位已不复存在。

2.2 区域差异显著

同时，地级市层面的银行业竞争格局也有显著的变化，东中部地区的银行业竞争程度逐渐增强，西部地区部分城市银行竞争程度变动较小。中国银行业垄断程度最大的地区集中在西藏、青海和新疆，这些地方的银行业金融机构类型较为单一。

参考姜付秀（2019）构建各个城市银行业的赫芬达尔-赫希曼指数 (HHI)，衡量各个省份的银行竞争水平。可以看到，北京、天津和上海的银行竞争程度最强，而青海、西藏和宁夏等西部城市的银行竞争程度较低，而且银行竞争强的地区和相对较弱的地区的差异明显。

图 6：各省份的银行竞争水平



资料来源：小组成员整理，中国银行保险监督管理委员会金融许可证信息数量库

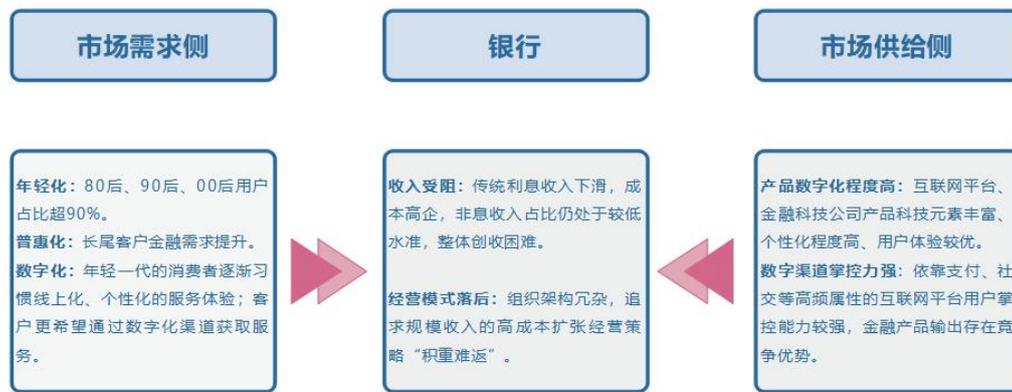
注：HHI 取值范围为 (0,1)，且当该指标越小时，表明银行竞争水平越高

2.3 数字化转型的破局效应

数字化转型能够提高商业银行的服务质量、降低服务成本，增强银行获客能力，因此商业银行在面临行业内部竞争压力时可能会采取数字化转型策略；除了行业内部竞争，银行业还需考虑来自行业外部如移动支付的冲击。

对此，我们可以从供需两端来进行考虑。从需求端来看，市场行为发生巨变，新一代年轻用户行为在向无纸化、移动化转变，对数字化渠道的熟悉程度也更高，更要求个性化服务。从供给端来看，金融科技公司、互联网平台等新市场参与者可以从供给端给消费者提供匹配需求的数字化产品及服务。

图 7：银行数字化转型供需端



资料来源：小组成员整理

3.数字化转型的量化分析

3.1 数字化转型的应用领域

当前银行业在数字化转型的过程当中，主要运用的核心底层技术包括：人工智能（AI）、大数据（Big Data）、云计算（Cloud）、区块链（Block chain），通过运用这四大新兴技术，使得银行在资源配置精准化、服务渠道全时化、业务流程自动化、风险管理智能化等方面提质增效显著。

表 1：数字化转型的技术与应用

底层技术	技术+金融	金融业应用的领域
人工智能	基于生物识别的人脸识别	账户远程开户、业务签约
	语音识别与处理	智能客服、营业网点机器人服务
	基于 OCR 自动化视觉处理	发票、合同、单据的信息结构化处理
	机器学习	智能投顾
	金融知识图谱	风险预警、反欺诈
大数据	客户画像	精准营销与获客

	大数据征信	授信评估、信用报告、贷中预警
云计算	云+大数据	海量数据云端存储
	拓展系统处理能力	满足系统高性能、容灾备份要求
区块链	物联网	确保底层资产交易真实可靠
	支付结算	支付收单机构间基于联盟链和智能合约实时自动对账

资料来源：小组成员整理

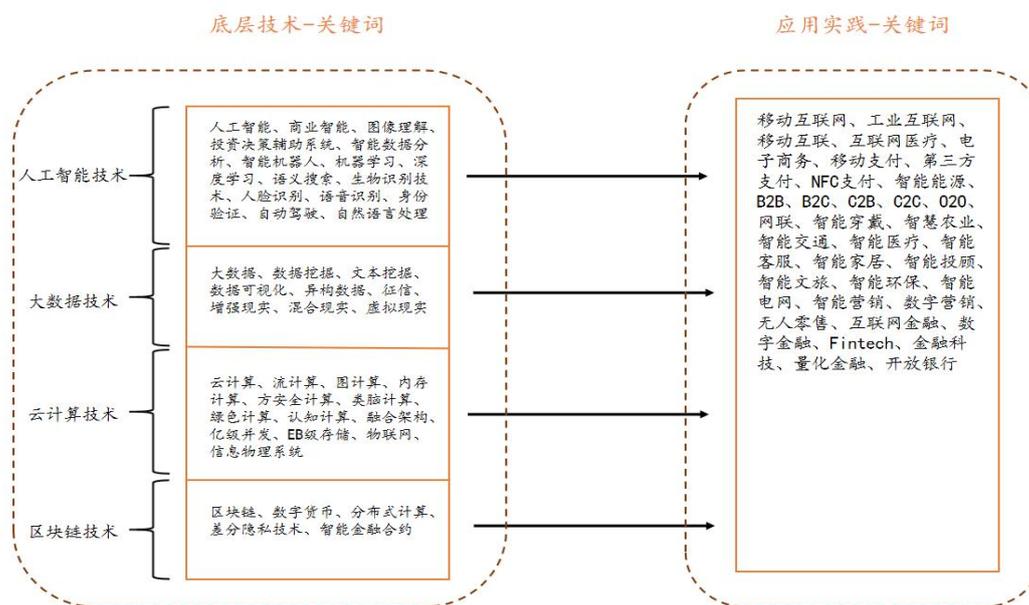
3.2 数字化转型的指标构造——基于文本分析

由于各个银行数字化转型的具体数据较难获得，而数字化转型这类特征信息做为发展决策更容易体现在企业的年/半年报文本中。因此，我们从上市企业（银行业）年/半年报中统计相关词频来刻画其转型程度。

我们通过 Python 爬虫功能归集整理了全部 A 股上市金融业的年/半年度报告，并通过 PDF 提取、分词等获得所有文本词汇，根据关键词进行搜索、匹配和词频计数，进而分类归集关键技术方向的词频并形成最终加总词频，从而得到企业数字化转型的关键度量指标。

在关键词的确定上，结合小组成员的讨论，并借鉴重要政策文件和研究报告的基础上，我们归纳整理出有关数字化转型的特定关键词如下：

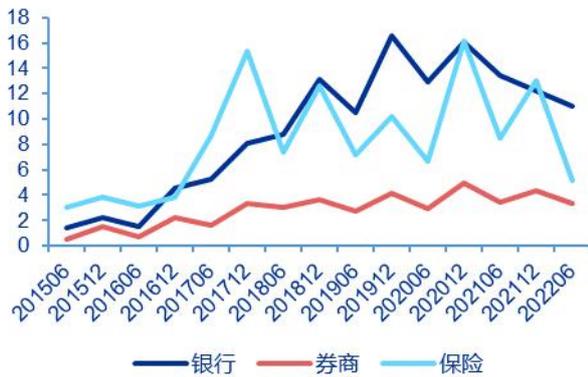
图 8：文本分析中的关键词说明



资料来源：小组成员整理

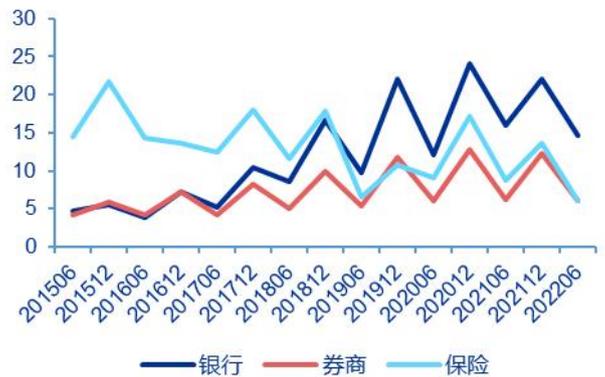
对关键词词频进行汇总统计后，我们发现 2016 年之前，保险业相比银行和券商行业，更加注重金融科技技术及其运用。随着 2017 年中国人民银行印发《金融科技发展规划（2019-2021）》开始，银行业开始加大对金融科技相关领域的投入，并于 2018 年超过保险业，成为金融科技技术及其运用的最重要的金融细分行业。

图 9：底层技术关键词词频（平均每家企业）



资料来源：小组成员整理

图 10：应用领域关键词词频（平均每家企业）



资料来源：小组成员整理

4. 银行业景气度的定量刻画

为刻画银行业景气历史变化，我们从供需角度出发，利用影响其主要利润来源的驱动因子来构建行业供需景气度。

对于银行业，供需结构主要是存贷款金额。供给的代理变量可以选择金融机构：各项存款余额；需求的代理变量可以选择金融机构：各项贷款余额；供需差异的代理指标可以选择金融机构：贷存差，其等于金融机构各项贷款余额-金融机构各项存款余额。

4.1 供需景气度指标筛选

4.1.1 表观供需差异

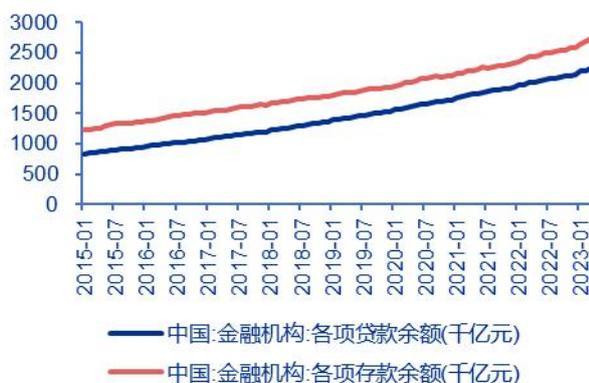
表观供给=金融机构各项存款余额

表观需求=金融机构各项贷款余额

表观供需差异=表观需求-表观供给=金融机构各项贷款余额-金融机构各项存款余额

选定金融机构贷存差（标准化数据）作为供需差异的代理变量。

图 11：金融机构各项贷款余额与存款余额



资料来源：Wind

图 12：金融机构贷存差与行业指数



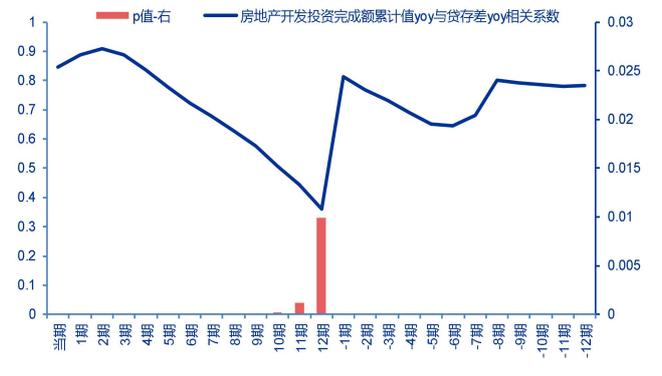
资料来源：Wind

4.1.2 潜在需求增速

基建和房地产是银行贷款的主要去向。故本文分别计算房地产开发投资完成额累计值（yoy）、固定资产投资完成额累计值（yoy）与贷存差（yoy）的相关性，以选取合适的潜在需求增速代理指标。

由相关性分析可知，房地产开发投资完成额累计值（yoy）与贷存差（yoy）呈显著正相关，且领先贷存差（yoy）2期；而固定资产投资完成额累计值（yoy）与贷存差（yoy）是负相关，故不予考虑。

图 13: 房地产开发投资完成额累计值 (yoy) 与贷存差 (yoy) 相关性分析



资料来源: Wind, 小组成员整理

图 14: 固定资产投资完成额累计值 (yoy) 与贷存差 (yoy) 相关性分析



资料来源: Wind, 小组成员整理

综上所述我们选择房地产开发投资完成额累计值（yoy）作为银行业潜在需求增速代理变量。

4.2 银行业供需景气度指标构造

选取两个指标作为供需景气度备选变量。

第一个指标是贷存差（yoy），其代表了银行业供需增速：

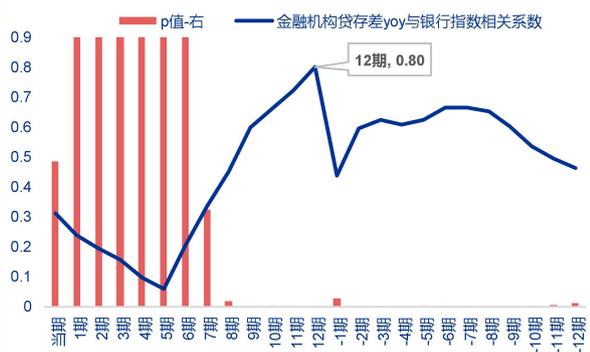
$$\text{金融机构贷存差 } yoy = Z(\text{金融机构贷存差 } yoy)$$

另一个指标采用以下公式构建而成：

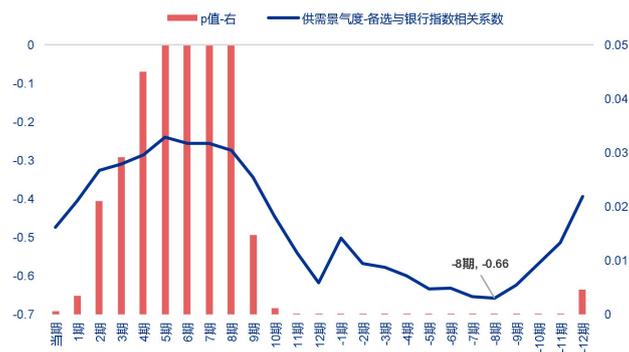
$$\begin{aligned} \text{供需景气度 - 备选} &= (\text{表观需求} - \text{表观供给}) * (1 + \text{潜在需求增速}) \\ &= Z(\text{金融机构贷存差}) * (1 + Z(\text{房地产开发投资完成额累计值 } yoy)) \end{aligned}$$

图 15: 金融机构贷存差 yoy 与行业指数相关性分析

图 16: 供需景气度-备选与行业指数相关性分析



资料来源：Wind，小组成员整理

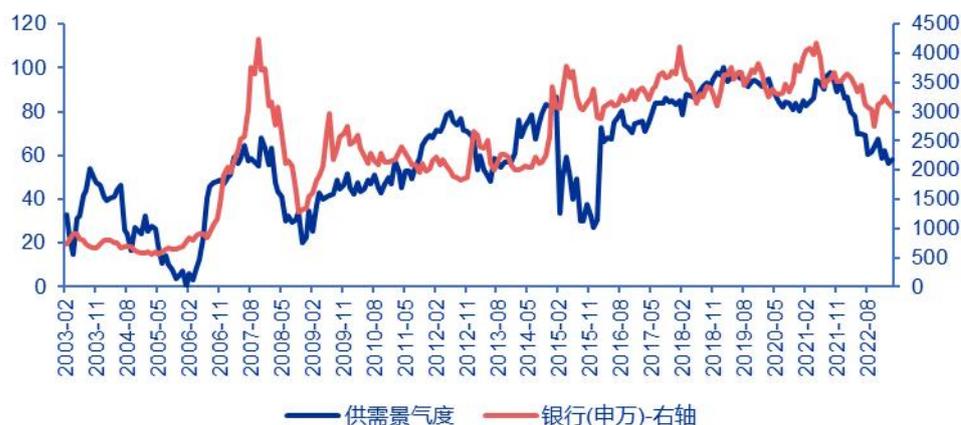


资料来源：Wind，小组成员整理

由相关性分析可知，金融机构存贷差（yoy）由于银行指数显著正相关，且领先银行指数 12 期（相关系数为 0.80）；而计算得出的供需景气度 2 与银行指数显著负相关，且滞后银行指数 8 期（相关系数-0.66）。

综上，选择领先效果和相关性均较好的金融机构存贷差 yoy 作为供需景气度指标。行业供需景气度（0-100 标准化处理）构造结果如下图所示：

图 17：行业景气度与申万银行指数



资料来源：Wind，小组成员整理

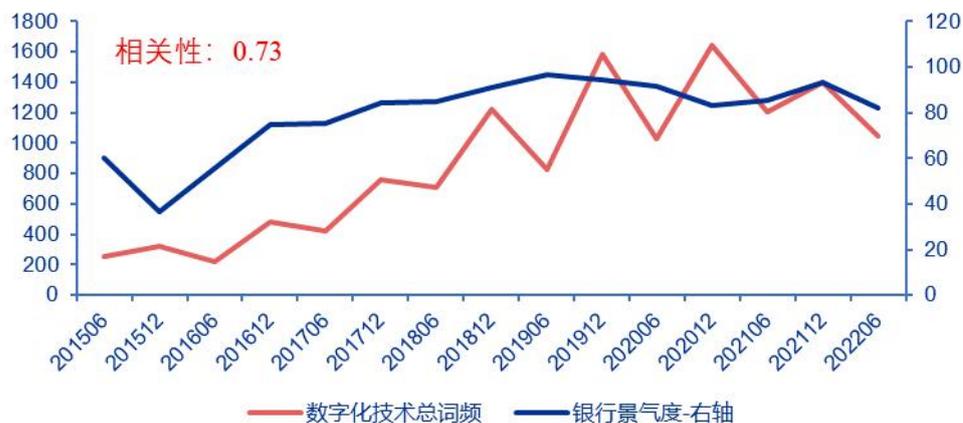
4.3 景气度表现

对比银行行业景气度（即行业供需景气度，以下简称行业景气度）与申万银行指数走势来看，我们刻画的景气度与指数走势较为一致，刻画效果较好。历史变化方面，银行行业在 2007 年迎来景气度的显著上升，同时申万银行指数也跃居高点，在随后的 2008 年下半年，受金融危机影响，银行景气度来到历史地位，在此后五年中复苏势头逐渐明朗。从近几年走势来看，景气度和银行指数均波动幅度较小，未出现明显拐点。

5.数字化转型与行业表现

因为构造的景气度为月频数据，为与半年报/年报频率相匹配，我们可以对景气度数据进行半年度降频处理。结合此前利用文本分析构建的数字化技术词频指标，我们观察发现其与银行行业景气度变化具有同步性，相关系数高达 0.73。

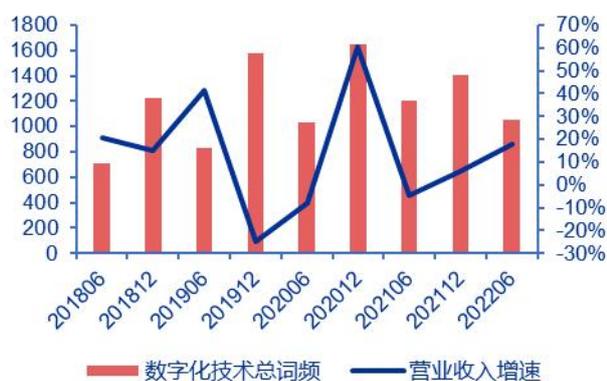
图 18：行业景气度与数字化技术总词频



资料来源：Wind，小组成员整理

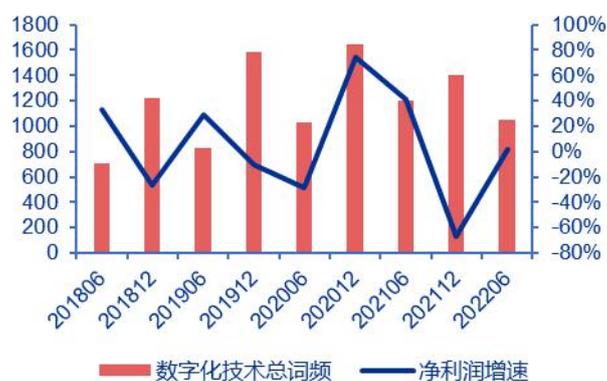
从上图可知，数字化技术总词频在 2019 下半年-2021 下半年维持在较高水平，并多次突破 1500 高点，直至 2022 上半年也维持在 1000 之上，2020 下半年总词频为 1646，为银行行业历史最高点。

图 19：数字化技术总词频与银行行业营业收入增速



资料来源：Wind，小组成员整理

图 20：数字化技术总词频与银行行业净利润增速



资料来源：Wind，小组成员整理

将营业收入增速/净利润增速与数字化技术总词频对比发现，两者与数字化技术总词频均具有一定的同步表现，尤其是在 2020 下半年，营业收入增速和净利润增速均位于历史最高位。

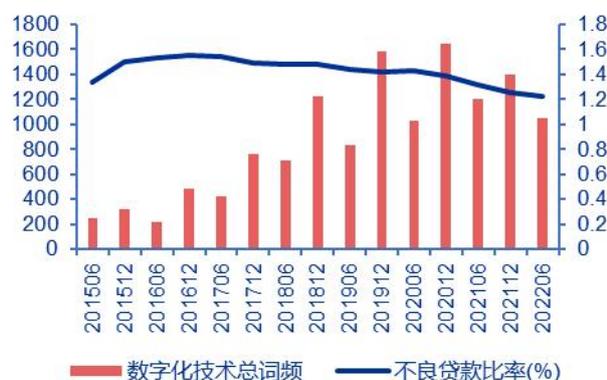
表 2：数字化转型各维度词频

日期	银行景气度	人工智能技术	大数据技术	区块链技术	云计算技术	数字技术运用	总词频
201506	60	3	45	0	9	196	253
201512	36	7	68	0	18	225	318
201606	55	3	47	1	13	155	219
201612	75	38	134	0	15	298	485
201706	75	38	137	26	13	212	426
201712	85	102	202	0	29	424	757
201806	85	100	171	53	36	349	709
201812	91	174	300	1	64	685	1224
201906	97	111	226	47	46	401	831
201912	94	212	368	2	99	900	1581
202006	92	123	269	82	55	499	1028
202012	83	184	362	11	101	988	1646
202106	85	125	251	105	71	653	1205
202112	93	133	265	14	88	903	1403
202206	82	88	223	83	57	597	1048

资料来源：小组成员整理

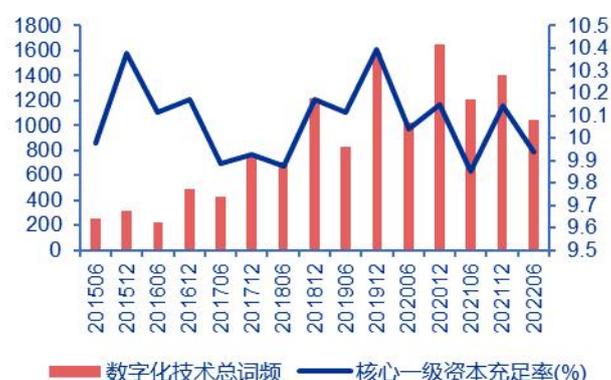
将总词频拆分到人工智能技术、大数据技术、区块链技术、云计算技术、数字技术运用五个维度，观察发现近几年有关大数据技术、数字技术运用两个维度词频较高，这两方面技术的发展或许可以帮助银行增强识别风险的能力。

图 21：数字化技术总词频与不良贷款率



资料来源：Wind，小组成员整理

图 22：数字化技术总词频与核心一级资本充足率



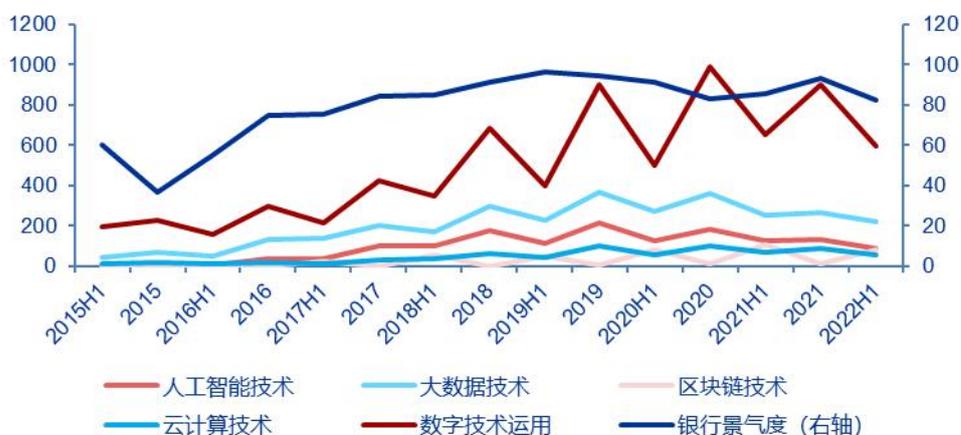
资料来源：Wind，小组成员整理

我国商业银行不良贷款比率呈逐渐走低的态势，与我国银行监管、风险防控的政策贯彻密不可分，因此难以捕捉到与数字化技术的关联度。从商业银行核心一级资本充足率来看，其与数字化技术总词频相关性较高，近几年相关性尤为显著。

因此我们有理由认为商业银行的数字化技术转型是有利于银行自身发展的，其一数字化技术可以利用数字平台和智能服务吸纳更多的客户以及提供更完善的服务，可以为自身带来更好的盈利能力及业绩表现；其二数字化技术可以显著提升商业银行自我风险监控、客户风险识别的能力，更好地应对相关政策的监管。

6.数字化转型下的行业展望

图 23：行业景气度与数字化转型五维度词频



资料来源：Wind，小组成员整理

上图给出了各类数字化技术词频与银行景气度之间的变动情况。如图所示，除区块链技术外，其他四种数字化技术词频呈现同样的变动趋势。其中银行景气度变化趋势与大数据技术词频以及人工智能技术词频之间的变动情况接近，而云计算技术词频与数字技术运用词频同银行业景气度之间变动情况的接近程度要稍弱一些。

表 3：数字化技术各维度与行业景气度相关性

数字化技术维度	人工智能技术	大数据技术	区块链技术	云计算技术	数字技术运用
景气度相关性	0.80	0.80	0.37	0.65	0.64

资料来源：小组成员整理

以上给出了各类数字化技术词频同银行景气度之间的相关系数，可以看出，人工智能技术词频与大数据技术词频同银行业景气度的相关性最高，达到了 0.80，云计算技术与数字技术运用词频相关系数次之，分别为 0.65 与 0.64，而区块链技术词频相关性最差，仅为 0.37。

综上所述，定性分析与定量结果均表明人工智能技术与大数据技术是考察银行数字化转型中需要关注的重点。

表 4：2021-2022 年人工智能技术词频与大数据技术词频 TOP5

年份	银行名称	股票代码	人工智能技术	大数据技术	合计
2022	邮储银行	601658	21	32	53
	平安银行	000001	9	26	35
	建设银行	601939	11	17	28
	中国银行	601988	10	9	19
	中信银行	601998	6	13	19
	小计			57	97

2022H1	邮储银行	601658	5	25	30
	上海银行	601229	13	12	25
	平安银行	000001	12	10	22
	中信银行	601998	5	16	21
	厦门银行	601187	1	15	16
	小计		36	78	114
2021	邮储银行	601658	7	29	36
	平安银行	000001	13	18	31
	渝农商行	601077	4	16	20
	上海银行	601229	14	5	19
	建设银行	601939	10	7	17
	小计		48	75	123
2021H1	邮储银行	601658	9	24	33
	平安银行	000001	12	17	29
	民生银行	600016	9	14	23
	上海银行	601229	13	9	22
	渝农商行	601077	5	15	20
	小计		48	79	127

资料来源: Wind, 小组成员整理

上表给出了 2021 年到 2022 年各期, 人工智能技术词频与大数据技术词频最高的五家银行, 从业绩来看, 邮储银行与平安银行人工智能技术词频与大数据技术词频一直在前五名, 上海银行有 3 期在前五名。

表 5: 邮储银行、平安银行、上海银行指标统计 (单位: 亿元、%)

银行名称	财务指标	2022	2022H1	2021	2021H1
邮储银行	总资产	140672.8	134264.2	125878.7	122170.5
	营业收入	3349.6	1734.6	3187.6	1576.5
	净利润	853.6	471.7	765.3	412.4
	不良贷款率	0.8	0.8	0.8	0.8
平安银行	总资产	53215.1	51087.8	49213.8	47235.2
	营业收入	1799.0	920.2	1693.8	846.8
	净利润	455.2	220.9	363.4	175.8
	不良贷款率	1.1	1.0	1.0	1.1
上海银行	总资产	28785.3	28146.5	26532.0	26283.6
	营业收入	531.1	279.4	562.3	277.1
	净利润	223.2	126.9	220.8	123.0
	不良贷款率	1.3	1.3	1.3	1.2

资料来源: Wind, 小组成员整理

由上表可以看出, 上海银行 2022 年人工智能技术词频与大数据技术词频没有进入前五名, 其营业收入同比下降, 净利润增长也不及邮储银行与平安银行。同时, 观察三家银行的

不良贷款率情况，邮储银行与平安银行均低于上海银行，其中邮储银行最低，各期均值为 0.83%。因此人工智能技术词频与大数据技术词频可以反映银行个股的业绩变化情况。报告以相同逻辑推荐投资者关注邮储银行与平安银行。

7.数字化转型风险提示

(1) 法律风险：金融科技兴起时间短，但发展迅速，法律的制定和通过存在一定的滞后性，使得金融科技的运用过程中存在较大的法律风险。

(2) 操作风险：金融科技的快速发展，对银行从业人员的专业性与技术性提出了更高的要求，内控或者员工培训的不完善很容易导致发生操作失误的现象。

(3) 信用风险：越来越多的银行在尽职调查和客户筛选过程方面大量应用数字化技术和新的模型算法，在大幅提高获客效率的同时，由于目前信用数据录入不完整，征信监管体系不完善等原因，导致信息披露不及时，信息对称的要求依旧没有达到，传统信用风险在金融科技技术的推动下，变得更加分散和隐蔽。

(4) 流动性风险：移动互联网平台上快速发展的创新金融产品，导致银行等机构资金来源稳定性下降。如部分银行通过蚂蚁金服、财付通等互联网平台出售其理财产品，在资金稳定性、客户粘性、收益敏感度等方面与传统资金来源存在显著差异，对银行的流动性管理构成新的挑战。

(5) 数据安全风险：在 21 世纪的今天，个人信息安全作为个人隐私的一部分，越来越受到人们的重视。而银行体系存在信息保管不善、内部员工泄露、遭受黑客攻击等问题，这类问题一旦发生，很容易使得该银行甚至整个行业的信誉蒙受损失。

(6) 技术风险：由于可能存在技术不成熟、算法缺陷与技术失控等风险，一旦数据与应用环境脱离可控范围，将导致金融科技市场失控，进而造成无法挽回的损失，甚至有可能对整个经济社会产生较为严重的负面影响。

(7) 道德风险：金融科技在银行业的应用，一定程度上降低了信息不对称，促进了行业的提质增效，但也产生了新型道德风险，如头部银行利用其行业影响力和雄厚的资金实力，大量招揽顶尖科技从业人员，规模效应导致行业集中度上升，降低相关服务的可替代性，进一步巩固大型金融机构“大而不能倒”的行业地位，增加金融体系中的道德风险。

附录

A.参考文献

[1]姜付秀,蔡文婧,蔡欣妮等.银行竞争的微观效应:来自融资约束的经验证据[J].经济研究,2019,54(06):72-88.

[2]吴非,胡慧芷,林慧妍,任晓怡.企业数字化转型与资本市场表现——来自股票流动性

的经验证据[J]. 管理世界, 2021, 37(07):130-144.

[3]金融科技是一把“双刃剑”，创新与风险并存. 中关村互联网金融研究院.

[4]中国金融科技生态白皮书. Caict.ac.cn. (2021). Retrieved 5 April 2022.

[5] 戚聿东，肖旭．数字经济时代的企业管理变革[J]．管理世界, 2020, 36(06):135-152+250. DOI:10.19744/j.cnki.11-1235/f.2020.0091.

B.金融科技子公司成立情况

表 6：金融科技子公司成立情况

母公司银行	金融科技子公司	成立时间	注册资本 (万)
央行	数字货币研究所	2017.01	2183
	深圳金科	2018.06	200
	长三角金科	2019.03	6820
	成方金科	2020.07	200780
	中汇金科	2020.01	10000
农业银行	农银金科	2020.07	60000
交通银行	交银金科	2020.01	60000
中国银行	中银金科	2019.06	60000
北京银行	北银金科	2019.05	5000
工商银行	工银金科	2019.03	60000
华夏银行	龙盈智达	2018.05	2100
民生银行	民生科技	2018.04	20000
建设银行	建信金科	2018.04	160000
光大银行	光大科技	2016.12	20000
招商银行	招银云创	2016.02	5000
平安银行	金融壹帐通	2015.12	120000
兴业银行	兴业数金	2015.11	50000
中信银行	中信金科	2018.05	11000
浦发银行	浦银金科	2021.05	100000
浙商银行	易企银	2020.02	2000
厦门国际银行	集友科技	2020.09	1000
廊坊银行	易达科技	2020.11	200
深圳农商银行	前海金信	2016.05	2100
浙江农信社	浙江农商数科	2020.12	10000
广西农信社	桂盛金科	2020.12	1200
盛京银行	盛银数科	2021.07	1000

资料来源：公开资料整理

C.银行数字化相关战略

表 7：银行数字化相关战略

	数字化战略	组织架构
工商银行	紧密围绕“十四五”时期金融科技发展规划和深化数字化转型总体方案推动“科技强行”和“数字工行”建设	构建运营新布局, 建立数据中台运营中心, 组建金融科技生态拓展团队以科技合作促进业务生态拓展
农业银行	深入推进信息科技 iABC 战略实施, 持续提升科技支撑和赋能水平	启动数字化业务运行中心、数据中心内蒙古分部建设, 建立科技与产品管理局、研发中心、数据中心
中国银行	深入推进集团全面数字化转型战略	成立金融数字化转型领导小组、金融数字化委员会, 强化产品创新与数字化发展的统筹融合与管理
建设银行	深化新金融行动, 发布《数字建行建设规划(2022-2025年)》, 发布数字力建设总体实施方案	首次界定“科技渠运板块”, 设“一办四部”: 数字化建设办公室、网络金融部、渠道与运营管理部、数据管理部、金融科技部
交通银行	将数字化转型作为战略实施的两大突破口之一, 聚焦价值创造和零售先行, 深化数字化、企业级思维, 以信息技术和数据要素双轮驱动, 建强数字化转型业务与技术平台支撑能力加快场景生态体系建设, 构建线上线下一体化经营新模式。	构建“两部三中心一研究院”: 金融科技部与数字化转型办公室、数据管理与应用部、测试中心、数据中心、软件开发中心公司、交银金科、金融科技创新研究院
邮政储蓄银行	将数字化转型作为全行转型发展的重点战略举措, 规划“123456”数字化转型战略布局, 推动“数字生态银行”建设	构建“三部两中心”: 信息科技管理部、金融科技创新部、数据管理部、软件研发中心、数据中心
招商银行	围绕线上化、数据化、智能化、平台化、生态化的目标, 从客户服务、风险管理、经营管理、内部运营等层面持续推进“数字招行”建设	设立“一部六中心”: 信息技术部、零售应用研发中心、批发应用研发中心、基础设施研发中心、数据资产与平台研发中心、测试中心、云数据中心

资料来源: 公开资料整理

D.网络爬虫及文本分析相关代码

```
import json
import os
import random
from time import sleep
from urllib import parse
import random
import requests
import pandas as pd
import os
import xlrd
```

```

from pdfminer.converter import PDFPageAggregator
from pdfminer.layout import *
from pdfminer.pdfparser import PDFParser
from pdfminer.pdfdocument import PDFDocument
from pdfminer.pdfpage import PDFPage, PDFTextExtractionNotAllowed
from pdfminer.pdfinterp import PDFResourceManager, PDFPageInterpreter
import xlwt
import os
import chardet

class crawl_wfs():
    '''
    2.5 特性:
    解决了*星问题
    增加了读取 excel(xlsx)的功能.xlrd模块需要为1.2版本.pip install xlrd==1.2
    新增了 error_list 记录未下载的错误股票代码
    '''
    def parsePDF(self, pdf_path, txt_path):
        # 以二进制读模式打开 pdf 文档
        fp = open(pdf_path, 'rb')
        # 用文件对象来创建一个 pdf 文档分析器
        parser = PDFParser(fp)
        # pdf 文档的对象, 与分析器连接起来
        doc = PDFDocument(parser=parser)
        parser.set_document(doc=doc)
        # 如果是加密 pdf, 则输入密码, 新版好像没有这个属性
        # doc._initialize_password()
        # 创建 pdf 资源管理器 来管理共享资源
        resource = PDFResourceManager()
        # 参数分析器
        laparam = LAParams()

        # 创建一个聚合器
        device = PDFPageAggregator(resource, laparams=laparam)
        # 创建 pdf 页面解释器
        interpreter = PDFPageInterpreter(resource, device)
        # 用来计数页面, 图片, 曲线, figure, 水平文本框等对象的数量
        num_page, num_image, num_curve, num_figure, num_TextBoxHorizontal = 0,
0, 0, 0, 0
        # 获取页面的集合
        for page in PDFPage.get_pages(fp):
            num_page += 1 # 页面增一
            # 使用页面解释器来读取
            interpreter.process_page(page)

```

```

        # 使用聚合器来获取内容
        layout = device.get_result()
        # 这里 layout 是一个 LTPage 对象 里面存放着 这个 page 解析出的各种
对象 一般包括 LTTextBox, LTFigure, LTImage, LTTextBoxHorizontal 等等 想要获取文
本就获得对象的 text 属性,
        for x in layout:
            if isinstance(x, LTImage): # 图片对象
                num_image += 1
            if isinstance(x, LTCurve): # 曲线对象
                num_curve += 1
            if isinstance(x, LTFigure): # figure 对象
                num_figure += 1
            if isinstance(x, LTTextBoxHorizontal): # 获取文本内容
                num_TextBoxHorizontal += 1 # 水平文本框对象增一
                # 保存文本内容
                with open(txt_path, 'a', encoding=self.encoding,
errors='ignore') as f:
                    results = x.get_text()
                    # print(results, end='')
                    f.write(results + '\n')
                print("正在解析第%s 页" % num_page)
                # print('对象数量: \n', '页面数: %s\n' % num_page, '图片数: %s\n' %
num_image, '曲线数: %s\n' % num_curve,
                # '水平文本框: %s\n' % num_TextBoxHorizontal)

def pdf2txt(self, ):
    files = os.listdir(self.folder)
    for file in files:
        if os.path.splitext(file)[-1] == ".pdf":
            self.pdf_name.append(file)
    for file_name in self.pdf_name:
        pdf_path = self.folder + "/" + file_name
        txt_path = self.folder + "/" + file_name[:-4] + '.txt'
        self.parsePDF(pdf_path, txt_path)

def get_adress(self, bank_name):
    url =
"http://www.cninfo.com.cn/new/information/topSearch/detailOfQuery"
    data = {
        'keyWord': bank_name,
        'maxSecNum': 10,
        'maxListNum': 5,
    }
    hd = {

```

```

        'Host': 'www.cninfo.com.cn',
        'Origin': 'http://www.cninfo.com.cn',
        'Pragma': 'no-cache',
        'Accept-Encoding': 'gzip, deflate',
        'Connection': 'keep-alive',
        'Content-Length': '70',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit
/ 537.36 (KHTML, like Gecko) Chrome / 75.0.3770.100 Safari / 537.36',
        'Content-Type': 'application/x-www-form-urlencoded;
charset=UTF-8',
        'Accept': 'application/json, text/plain, */*',
        'Accept-Language': 'zh-CN, zh; q=0.9, en; q=0.8',
    }
    r = requests.post(url, headers=hd, data=data)
    # print(r.text)
    r = r.content
    m = str(r, encoding="utf-8")
    pk = json.loads(m)
    try:
        orgId = pk["keyBoardList"][0]["orgId"] # 获取参数
        plate = pk["keyBoardList"][0]["plate"]
        code = pk["keyBoardList"][0]["code"]
        zwjc = pk["keyBoardList"][0]["zwjc"]
        # print(orgId, plate, code)
        return orgId, plate, code, zwjc
    except:
        return 'a', 'b', 'c', 'd'

def download_PDF(self, url, file_name): # 下载 pdf
    url = url
    # 遇到*号替换掉
    if "*" in file_name:
        file_name = file_name.replace("*", '')
    r = requests.get(url)
    #创建文件（文件命名不能有*号）
    f = open(self.folder + "/" + file_name + ".pdf", "wb")
    f.write(r.content)

def get_PDF(self, orgId, plate, code, zwjc):
    url = "http://www.cninfo.com.cn/new/hisAnnouncement/query"
    data = {
        'stock': '{} {}'.format(code, orgId),
        'tabName': 'fulltext',
        'pageSize': 30,
    }

```

```

        'pageNum': 1,
        'column': plate,
        'category': 'category_bndbg_szsh;',
        'plate': '',
        'seDate': self.seDate, # 设置时间
        'searchkey': self.searchkey,
        'secid': '',
        'sortName': '',
        'sortType': '',
        'isHLtitle': 'true',
    }

    hd = {
        'Host': 'www.cninfo.com.cn',
        'Origin': 'http://www.cninfo.com.cn',
        'Pragma': 'no-cache',
        'Accept-Encoding': 'gzip, deflate',
        'Connection': 'keep-alive',
        # 'Content-Length': '216',
        'User-Agent': 'User-Agent:Mozilla/5.0 (Windows; U; Windows NT 6.1;
en-US) AppleWebKit/533.20.25 (KHTML, like Gecko) Version/5.0.4 Safari/533.20.27',
        'Content-Type': 'application/x-www-form-urlencoded;
charset=utf-8',
        'Accept': 'application/json, text/plain, */*',
        'Accept-Language': 'zh-CN, zh;q=0.9, en;q=0.8',
        'X-Requested-With': 'XMLHttpRequest',
        # 'Cookie': cookies
    }

    data = parse.urlencode(data)
    # print(data)
    r = requests.post(url, headers=hd, data=data)
    # print(r.text)
    r = str(r.content, encoding="utf-8")
    r = json.loads(r)

    reports_list = r['announcements']
    try:
        for report in reports_list:
            # print(report)
            # if '摘要' in report['announcementTitle'] or "20" not in
report['announcementTitle']:
                # continue
            if '摘要' in report['announcementTitle']:
                continue

```

```

        if '正文' in report['announcementTitle']:
            continue
        if 'H' in report['announcementTitle']:
            continue
        else:
            #
            http://static.cninfo.com.cn/finalpage/2019-03-29/1205958883.PDF
            pdf_url = "http://static.cninfo.com.cn/" +
report['adjunctUrl']
            # 社会责任报告，中间会有特殊字符，替换掉 2017<em>社会
</em><em>责任</em>报告
            my_str = report['announcementTitle'].replace('<em>',
''').replace('</em>', '')

            file_name = "{}-{}-{}".format(my_str, code, zwjc)
            print("正在下载: " + pdf_url, "存放在当前目录: /" +
self.folder + "/" + file_name)
            self.download_PDF(pdf_url, file_name) # 主要功能，下载
pdf

            sleep(random.random())
    except:
        self.error_list.append([code, zwjc])

# 加载 txt 列表寻找关键词并保存到 excel
def matchKeyWords(self, txt_folder, keyWords):
    files = os.listdir(txt_folder)
    words_num = [] # 保存所有文件词频
    for file in files:
        word_freq = {} # 单词出现频率次: word: num
        if os.path.splitext(file)[-1] == ".txt":
            txt_path = os.path.join(txt_folder, file)
            with open(txt_path, "r", encoding=self.encoding,
errors='ignore') as fp:
                text = fp.readlines()
                for word in keyWords:
                    num = 0
                    for line in text:
                        num += line.count(word)
                    word_freq[word] = num
                stock_code = file.split("-")[1]
                stock_name = file.split("-")[2]
                year = file.split("-")[0][0:4]
                words_num.append((word_freq, stock_code, stock_name,
year))

```

```

book = xlwt.Workbook(encoding='utf-8', style_compression=0)

sheet = book.add_sheet('年报关键词词频统计', cell_overwrite_ok=True)
sheet.write(0, 0, '年份')
sheet.write(0, 1, '企业代码')

for i in range(0, len(keyWords)):
    sheet.write(0, i + 2, keyWords[i])

for index, one in enumerate(words_num):
    word_f = one[0]
    stock_code = one[1]
    stock_name = one[2]
    year = one[3]
    for ind, word in enumerate(keyWords):
        sheet.write(index + 1, ind + 2, word_f[word])
    sheet.write(index + 1, 0, year)
    sheet.write(index + 1, 1, stock_code)
    # sheet.write(index + 1, 1, stock_name)

book.save(self.folder + '\年报关键词词频统计.xls')

def matchKeyWords2(self, txt_folder, keyWords):
    files = os.listdir(txt_folder)
    words_num = [] # 保存所有文件词频
    for file in files:
        word_freq = {} # 单词出现频率次: word: num
        if os.path.splitext(file)[-1] == ".txt":
            txt_path = os.path.join(txt_folder, file)
            with open(txt_path, "r", encoding=self.encoding,
errors='ignore') as fp:
                text = fp.readlines()
                alltext = ''
                for line in text:
                    alltext += line.replace("\n", "")
                for word in keyWords:
                    num = 0
                    num += alltext.count(word)
                    word_freq[word] = num
            stock_code = file.split("-")[1]
            stock_name = file.split("-")[2]
            year = file.split("-")[0][0:4]
            words_num.append((word_freq, stock_code, stock_name,
year))

```

```

book = xlwt.Workbook(encoding='utf-8', style_compression=0)

sheet = book.add_sheet('年报关键词词频统计', cell_overwrite_ok=True)
sheet.write(0, 0, '年份')
sheet.write(0, 1, '企业代码')
# 自定义表格
for i in range(0, len(keyWords)):
    sheet.write(0, i + 2, keyWords[i])

for index, one in enumerate(words_num):
    word_f = one[0]
    stock_code = one[1]
    stock_name = one[2]
    year = one[3]
    for ind, word in enumerate(keyWords):
        sheet.write(index + 1, ind + 2, word_f[word])
    sheet.write(index + 1, 0, year)
    sheet.write(index + 1, 1, stock_code)
    # sheet.write(index + 1, 1, stock_name)

book.save(self.folder + '\年报关键词词频统计.xls')

def judgmentTextEncoding(self, txt_folder, ):
    files = os.listdir(txt_folder)
    for file in files:
        if os.path.splitext(file)[-1] == ".txt":
            txt_path = os.path.join(txt_folder, file)
            # print(txt_path)
            break
    with open(txt_path, 'rb') as f:
        text = f.read()
        info = chardet.detect(text)
        # print(type(info))
        self.encoding = info['encoding']
        # print(self.encoding)

def __init__(self, bank_list, seDate, kw):
    self.folder = ''
    self.category = ''
    self.bank_list = bank_list
    self.seDate = seDate
    self.kw = kw.split(",") # 这里方
    self.pdf_name = []

```

```

self.encoding = "utf-8"
self.searchkey = ""
self.error_list = []

def modechange(self, num):
    if num == 1:
        self.folder = '半年报'
        self.category = 'category_bndbg_szsh'
    elif num == 2:
        self.folder = '年报'
        self.category = 'category_ndbg_szsh'
    elif num == 3:
        self.folder = '一季报'
        self.category = 'category_yjdbg_szsh'
    elif num == 4:
        self.folder = '三季报'
        self.category = 'category_sjdbg_szsh'
    elif num == 5:
        self.folder = '社会责任'
        self.category = ''
        self.searchkey = '社会责任;'
    elif num == 0:
        self.folder = '自定义'
        self.category = '' #目录分类为空
        self.searchkey = '社会责任;' #这里自定义, 搜索的关键词

def step1(self):
    '''爬取 pdf'''
    if not os.path.exists(self.folder):
        os.mkdir(self.folder)
    # 爬取 pdf 并转换为 txt
    for bank in self.bank_list:
        orgId, plate, code, zwjc = self.get_adress(bank)
        self.get_PDF(orgId, plate, code, zwjc)
    print("pdf All done!")
    print(self.error_list)

def step2(self):
    '''pdf2txt'''
    print('开始 pdf2txt!!!!')
    self.pdf2txt()
    print("pdf2txt finish")

```

```

def step2_5(self):
    print('开始提取文本编码格式!!!')
    self.judgmentTextEncoding(self.folder)

def step3(self):
    '''词频统计'''
    print('开始统计词频!!!!')
    # self.matchKeyWords(self.folder, self.kw)#不能统计到换行关键字
    test.matchKeyWords2(self.folder, self.kw) # 可以统计到换行关键字,
可能有未知 bug
    print('统计结束')

def bank_xlsx(xlsxPath):
    data = xlrd.open_workbook(xlsxPath)
    table = data.sheets()[0]
    coll_value = table.col_values(0)
    return coll_value

if __name__ == '__main__':
    # bank_list = ['000001', '000002', '000003', '000004', '000005']
    bank_list = [...]
    seDate = '2015-05-31~2016-05-31' # 时间设置
    kw = "人工智能、商业智能、……" # 关键词, 顿号隔开

    test = crawl_wfs(bank_list, seDate, kw) # 实例化
    test.modechange(1) # 这里选择模式
    test.step1() # get pdf, 如果不下载文件, 跳转到 get_PDF 查看下载行是否被
注释
    test.step2() # pdf2txt
    test.step2_5() # 判断文本编码格式
    test.step3() # 词频统计

```