ANNOTATION GUIDELINE

Mingyuan Ma, Shuyao Zhou, Yuanrui Zhu

General Principle:

- **Pipeline of classes:** input document → BLOCK1 → BLOCK2 → RESIDUAL

  BLOCK1 is used to determine if a document is class-1 or 3 or 4 based on the strong or multiple weak signs of class-1/3/4; if it can be determined, then output the class; else, move on.
  Output: class-1 or class-3 or class-4
  BLOCK2 is used to determine if a document is class-2, which is between class-1 and class-3, based on the signs of class-2. Generally, if a document is class-4, it would possess obvious signs/features, majority of what could be determine in BLOCK1
  Output: class-2 or class-1
  RESIDUAL is used to determine those documents that cannot be classified by BLOCK1 and BLOCK2. In this layer, subjective judgments are involved. In general, we need to consider the sentence holistically, and consider which kind of signs are dominated in quantity and intensity. Hence, the output would usually be class-2 and class-3. For consistency, we would not set the output to be class-1. And we would discuss it later.
  Output: class-2 or class-3
  In any blocks, if one document is output as one class, then it would not be transferred to its next block (keep moving on). This point is important in our logic of annotation.

- **How to classify:** First, read the sentence and check those characteristics/signs of class-1 to 4 and their intensity (weak/strong). Second, pay attention to caveats and “what does not count as a sign”. Third, consider holistically based on the descriptions of four classes.

- **Before annotating:** We have removed all special html symbols and links, so we are only focusing on their plain texts, characters, and punctuations.

- **If one document contains quotations, remove the quotations before analyzing:**

  Examples:
  Complete document: I don't remember who said it, but I vaguely remember there's a quote amounting to "You are never the same person you where before reading a book".

What we consider: I don't remember who said it, but I vaguely remember there's a quote amounting to [quote]

- **If one document contains special html symbols, remove them.**
  Examples**:**
  Complete document: >To me, for magic to feel mystical it would need to instead feel subtler rather than more chaotic/unreliable/fickle, and/or involve a heavy degree of ritual to it.
  *What we consider: T*o me, for magic to feel mystical it would need to instead feel subtler rather than more chaotic/unreliable/fickle, and/or involve a heavy degree of ritual to it.

- **If one document contains numerous references or web links, we use [outside link] or [link] to indicate whether it refers to an outside source.**
  Examples:
  Complete document: Electoralism is a term first used by Terry Karl, professor of political science at Stanford University, to describe a "half-way" transition from authoritarian rule toward democratic rule. As a topic in the dominant party system political science literature, electoralism describes a situation where the transition out of hard-authoritarian rule is initiated and managed by the incumbent regime. More details here: https://en.wikipedia.org/wiki/Electoralism
  What we consider: Electoralism is a term first used by Terry Karl, professor of political science at Stanford University, to describe a "half-way" transition from authoritarian rule toward democratic rule. As a topic in the dominant party system political science literature, electoralism describes a situation where the transition out of hard-authoritarian rule is initiated and managed by the incumbent regime. More details here: [outside link]

- **If the document is automatically generated by the system or is a bot, we would analyze this document from other signs of it, rather than classify it directly to one of those classes.**

Characteristics of casual expression in annotation (class-1):

Before we start: We cannot quantify the intensity of Weak and Strong of each sign since we all know that language is subjective. But we would approximately regard, in 0~100 scale, Weak would be in 10-35, while Strong would be 65-100. We leave the gap intentionally to distinguish between those two classes and give less degree of freedom when evaluating and less disagreements. Again, we totally understand it is hard to quantify, it is subjective.

A. Weak signs of casual expression:

- **Imperative Sentence or sentence start with a verb**
    a. Can’t remember the author or guests name though
    b. Love the movies but would quite like to see a more faithful adaptation

- **Capital letter to empathize something**
    a. You shouldn’t shun EVERYTHING that becomes trendy just because it’s trendy.

- **Consists of short/colloquial rhetorical question(s) or general question(s)**
    a. What's titel? Does it work for non-English books too?
    b. “What’s the meaning of life? any definition?
    c. r/books gave you permission to post this on r/AskLiteraryStudies?

- **Very short colloquial sentences that express personal feelings**
    a. Such is life.
    b. Up to you.
    c. What a book!

- **Frequent use of first person as the subject of sentences (almost every sentence)**
    a. It's a florist at the moment but I'd love to one day open it up as a little used book shop. I'm sure I would never make any money but I could read and talk about books if anyone actually came up the street to browse in it.

- **Minor misuse of grammar**
    a. But not just in the most obvious indignities for people of color.

- **Contain absolute words with a not formal (may be casual or ordinary) adjective**
    a. It is definitely interesting.
    b. The book is extremely boring

  Comment: Of course, those absolute words are not a necessary sufficient sign of casual expression, so we would regard it as a weak sign of casual language. However, in most cases, we may need to determine if these absolute words to be a sign based on the adjective or noun after it. For example, “the book is definitely nothing” would be

B. Strong signs of casual expression:

- **Contain emoji or using non-English punctuations/marks to represent words**
    a. The $ goes to breast cancer research.
    b. I haven't used it before :(
    c. I ever be the same human again 😂

Comment: $75k is totally fine since there "$" is not to replace a word but as a unit.

- **Contain colloquial words or phrase or slang or short**
    a. I would bet good money that you won't read a book about it.
    b. legit, damndest, -ish
    c. use "cause" instead of "because"; use "though" instead of "although"
    d. Pretty stupid
    e. Stuffs like that…

- **Start with a *non-sentence* follow up**
    a. **"**Same", "Same here"
    b. "No problem", "True"
    c. "Agree", "Disagree"

    Comment: If one starts with a *sentence (SVO) follow up*, then it would not be a sign of casual expression. For example, "I agree with what you mention."

- **Contain internet acronyms**
    a. "TBH" is the internet acronyms of "to be honest"
    b. "ASAP" is the internet acronyms of "as soon as possible"
    c. "LOL" is the internet acronyms of "laughing out loud"

- **An obvious expression including exclamation of surprise, wonder, pleasure, or the like, which we can observe from *words* and *punctuations***
    a. Oh wow, he does have a lot of books! I'll see which ones my library carries and go from there. Thank you!

- **Use "…" to represent speechless**
    a. I feel like we'd be similar people...though these are sweeping generalizations.
    b. Pretend to ignore all of that…

    Comment: when "..." is used to enumerate something, then it would not be regarded as a weak sign of casual expression. For example, "sunshine, beach, waves…" is not a sign of casualness but a sign of artistic expression.

- **If the comment is an incomplete sentence.**
    a. Agree
    b. Great
    c. r/suggestmeabook.
    d. An X-Men movie?

C. What does not count as a sign of casualness:

- **We are not considering those non-English characters which have no literal meaning. For example, “.” “?” “,”**

- **Some abbreviations would not be considered as a sign due to high frequency of usage. BUT only limit to those below, ALL other abbreviations would be regarded as *weak signs of casual expression.***
    a. I’m == I am; I’ll == I will; I’d == I would
    b. / == or
    c. didn’t == did not; don’t == do not; can’t == cannot; hadn’t == had not; haven’t == have not; hasn’t == has not

  Comment: For example, “We’d” would be regarded as a weak sign of casualness due to the low frequency of usage in our exploration data. “GoO”, “S3s” would be regarded as a weak sign of casualness as well by assumption even it may truly be the name of something, but we do such assumption for all data and it is just a weak sign not a strong sign, so it may not affect the evaluation of a document too much.

- **Abbreviations of proper nouns or phrases shouldn’t be automatically categorized as informal (i.e. PVE). Other aspects of the answer should be carefully considered.**

- **If the writer uses the full name of those abbreviations above, then it would be considered as *weak signs of formal expression.***

Characteristics of ordinary expression in annotation (class-2):

Before we start:

“Trick” for class-2: if you find it could not be formal (class-3), and you cannot initially or instantaneously figure out if it should be class-1 (casual), then you can label it as class-2. The reason is because if you cannot quickly classify it to 1, then there may lack strong signs of casual expression, then it should be 2.

However, “it could not be formal (class-3)” refers to it would 100% not be class-3 (we are there to discuss the boundary between c1 and c2). And this trick or boundary cannot be used to distinguish class-2 and class-3! For the boundary between c2 and c3, please check RESIDUAL

which we would elaborate below! And again such boundaries between two classes are subjective, but what we write down there is what the majority of documents follow.

And we use “trick” there rather than a guideline because such boundaries are so subjective that they are hard to quantify by conditions/rules/guidelines with 100% accuracy. There would always exist edge cases, but we would clarify as much as possible below.

- **If the document does not satisfy any condition of class-1 (casual) or any condition of class-3 (formal) or any condition of class-4 (artistic), then would be regarded as class-2 (ordinary)**
    a. I don't remember who said it, but I vaguely remember there's a quote amounting to [quote].

- **If the document contains only weak sign of casual expression and cannot be formal (class-3) or artistics (class-4), then it would be considered as class-2 (ordinary)**

- **If the document has no signs (weak/strong) of formal or casual expression (eg. you cannot directly tell its class), and the point of view is “I”, and if there is NO evidence or reason to bolster a document's opinion or expression**
    a. Also, this is truelit, I come here for in-depth literature discussion, not to be reminded of reddits tendency to review books they don't plan on reading.
    Comment: if there is ANY evidence/reason, then it would be class-3

- **If contain neglectable (weak) signs of class-1/3/4 compared with the whole sentence, then it would be regarded as class-2 (ordinary)**

    a. Whether that's a good thing or not probably depends on your perspective and preconceptions of the books/authors you read and the lessons learned from them.
    Comment: Although we can observe this document use “/” to refer to “or,” it is just a weak sign of casual expression. Compared to the whole sentence, it is neglectable. Hence, it would be classify to class-2 rather than class-1

- **If the document does not belong to class-4 (artistic), and it contains approximately equal total intensity/tendency of signs in class-1 (casual) and class-3 (formal), then would be regarded as class-2 (ordinary). There we can assume “ambiguity” is a way of “ordinary.”**

- **If the document is just consists of sentence(s) which does not contain any obvious signs of casualness, then it would be considered as class-2**
    a. That flour yeast water book?
    b. The only book you will need
    c. I know the one has allegedly been around for a while. And I have no idea what

other information he gives out. But from what I saw on that post talking about the importance of emotional connection to your dick and some of the comments he was making made me cringe. And on top of that the fact that he claimed to gain so much in so little I just instantly would not believe anybody.

- **If the document just provides person expression or suggestion, without any further discussion or explanation, it should be classified as class-2**

Characteristics of formal expression in annotation (class-3):

Before we start: It is harder to determine if a document is formal than it is casual since we can have weak/strong signs of casualness, while to classify a document as formal we cannot find a single rule that is more indicative than others. Even though we can have one, a single strong sign of casualness may lead the document to become casual or indeterminate. Hence, unlike casual expressions where we have Weak/Strong signs, there for formal expressions we have an assumption that those signs are equally important.

- **If the document is only one sentence to direct the reader to a different post (with no signs of casual expression), or a suggestion to different post, then it would be formal**
    a. Please post lists of books in /r/booklists.
    b. They'll be in r/truelit in no time.
    c. As far as I know, recommendation requests should go [outside link].
    Comment: The reason is that there is no strong sign of class-1, then we should see if there are any weak signs of class-2. Since there is no weak sign of class-2, we should see if there are any signs of class-4. Because there is no sign of class-4, then it is classified as class-1.

- **If the document is just describing a scene from literature, then it would be classified as formal. The reason is that it may be used as an evidence or literature technique to prove the style of the author. And we need to make sure of the consistency.**
    a. Kate Daniels is a down-on-her-luck mercenary who makes her living cleaning up these magical problems.
    b. The Masters of the Dead, necromancers who can control vampires, and the Pack, a paramilitary clan of shapechangers, blame each other for a series of bizarre killings.

- **If the document is just doing the literature review with non-colloquial wording or quoting facts in literature *or describing a scene in literature*, then it would be classified as formal (similar to the first sign but may different slightly)**
    a. In 'Destined to Witness', Hans Massaquoi has crafted a beautifully rendered memoir -- an astonishing true tale of how he came of age as a black child in Nazi Germany.
    b. Alek and Deryn are aboard the Leviathan when the ship is ordered to pick up an unusual passenger.

- **The level of vocabulary used in the document. If the words used in one document are more sophisticated, more professional, and more formal, then it would be a sign of formality. Since it may be more subjective to define, it would just be regarded as a Weak sign**
    a. how the military justice system operates." *would sound more formal than* "how the military justice system works.

- **The document is objectively describe a fact with formal (not colloquial) wording without any opinion**
    a. At the heart of his theory, known as logotherapy, is a conviction that the primary human drive is not pleasure but the pursuit of what we find meaningful.
    b. In My Stroke of Insight, the author shares her recommendations for recovery and the insight she gained into the unique functions of the two halves of her brain.
    c. The Santa Fe Institute (SFI) is an independent, nonprofit theoretical research institute located in Santa Fe, New Mexico [...] computational, biological, and social systems.

- **The document describes a fact with formal (not colloquial) wording as evidence to prove some opinions.**

    a. Joseph Smith was extremely anti-Christian due to the fact that they laughed the loudest at his nonsense [...]

- **The document describes a fact from the view of the author or a character**
    a. In an attempt to be unique and separate from them, Joseph thought they were not Christian.
    b. Psychiatrist Viktor Frankl's memoir has riveted generations of readers with its descriptions of life in Nazi death camps and its lessons for spiritual survival.

- **The tone is serious and official rather than those extreme expressions, including lighting, friendly, cursing, etc. However, it does NOT mean the document could NOT contain expressions inside.**

    a. There are a lot of procedural and ethical issues with how the military justice system operates. Specifically, the civilian system allows someone the ability to

have those things removed permanently and they are only collected after an arrest which only happens with probable cause.

- **Use factual evidence or personal fact, but such personal fact should not contain affection or tendency**

    a. The anthropologist shows that for more than 5,000 years, since the beginnings of the first agrarian empires, humans have used elaborate credit systems to buy and sell goods

Characteristics of artistic expression in annotation (class-4):

Before we start:

The type of artistic expression is unique in our annotation. It has some clear characteristics that distinguish it from other classes.

- **Contain rhythm inside and the wording is not colloquial or formal**
    a. Diaries might be just silly enough to make you smile; Sending you all my sympathy for your sad loss.
    b. True love or not, the game must play out, and the fates of everyone involved, from the cast of extraordinary circus performers to the patrons, hang in the balance, suspended as precariously as the daring acrobats overhead.
- **Contain stylic words or employ symbolism in the sentence**
    a. When the magic is up, rogue mages cast their spells and monsters appear, while guns refuse to fire and cars fail to start. But then technology returns, and the magic recedes as unpredictably as it arose, leaving all kinds of paranormal problems in its wake.
    b. The plan works like a charm—at first. But amid the glittering, gossipy, cut-throat world of London's elite, there is only one certainty: love ignores every rule.

    Comment: if the document itself is related to something that is magic, then it would not become a sign of class-4

- **Contains a stylistic description (figurative language) to bring readers in**
    a. In the dark forest at the end of the road, lies something buried. But, it is not what you think.
    b. Your life will slowly evaporate elsewhere until you cease to exist in the form you label as yourself.

- **The structure of the sentence is not grammatically correct, but is stylistic**

a. “Your time, quality time together; a day at the spa; a weekend away; Stuff she likes, maybe novels/books; maybe wine; coffee gift cards; restaurant gift cards; a new pet; jewelry; tickets to events that she's into;”

What cannot be regarded as artistic:

- **If majority of sentences in document is belongs to class 1/2/3, even there is one stylistic sentence, the document cannot be regarded as artistic**
    a. Literature suffers from the same problems. Truelit suffers from completely different problems namely lack of threads, small scope, and lack of self awareness -- though being under-read and their identifying as readers can also be an issue. 52books as well. *The grass is not greener elsewhere.*

Summary and How to make decisions in RESIDUAL BLOCK

Before we start: as what we mention above of three different blocks we need to consider throughout our annotation process. BLOCK1, BLOCK2, and RESIDUAL. All signs we mention above are needed for our judgment in all blocks. Below, are signs that may help us to determine in the residual block, where it filters those documents which do not contain obvious signs that can be determined in BLOCK 1/2.

- We understand that it would be hard to distinguish between class-2 and class-3 especially those documents on the “boundary.” We acknowledge that all the rules in this guideline cannot be generalized to all cases in this word, and our interpretation of language is subjective, which cannot be quantified consistently to everyone with different background, including native language, education, or even the mood of the day when you are labeling those documents. However, we have to make some guidelines to ensure the consistency of our judgments as much as possible. And all members in our group discuss those rules for a long time to ensure those “rules” are reasonable.

NOTICE that we only need to consider RESIDUAL BLOCK after BLOCK 1 and BLOCK 2, so we cannot use the signs below to judge when we are in BLOCK1 or BLOCK 2.

- **Any suggestion with NO obvious some of casual expression would be regarded as either class-2 or class-3 (cannot be class-1), and if there contains sign of casualness then would be class-2, otherwise would be class-3**
    a. “r/AskLiteraryStudies is probably worth checking out if you want more in-depth answers on job opportunities” will be class-2 since it contains “checking out”.
    b. “Check r/AskLiteraryStudies.” will be class-2 since it is an imperative sentence whis is a sign of casualness.

c. “r/AskLiteraryStudies is helpful.” will be class-3 since although there is no signs of formality, there is no sign of class-2

Comment: however, if there exist obvious signs of casual expression (lexical, structural, etc.), we would classify it as class-1. Why? Because remember we are annotating through a pipeline. This kind of document cannot be in the RESIDUAL block because it would be labeled/output in BLOCK 1.

- **If the document contains obvious *personal* affection or tendency, but with NO evidence/reason to support and the adjective/verb is NOT advanced, then it would be class-2:**
    a. I do not like this book
    b. I hate you
    c. I like it

- **If the document contains obvious *personal* affection or tendency, but with evidence/reason to support AND such evidence/reason should not contain personal affection or tendency or opinion, then it would be class-3:**
    a. I do not like this book because it contains ideas againsting our common sense.
    b. I agree with these statements. [reason1, reason2, reason3, etc.]

  Comment: “I do not like this book because it is boring” should still be class-2 since the reason (it is boring) is still containing personal affection. However, “I do not like this book because it contains ideas againsting our common sense.” is class-3 since the reason (it contains ideas againsting our common sense) is a factual evidence with no personal affection/opinion inside

- **If the document contains obvious *personal* affection or tendency, but with the adjective/verb is advanced, then it would be class-3:**
    a. This antique architecture is magnificent
    b. I disdain it

- **If the document refers to [outside link] or using factual evidence, then we regard this as an indicator of supporting material. Or even giving evidence after the starting senetcen that is agreement/disagreement So, the document will be classified as class-3.**
    a. I'm going to have to disagree with you. [reason1, reason2, reason3, etc.]

- **If the document does not contain any personal information including opinion or affection or tendency, then it would be classified as class-3**