The outline that you submit/write-up to Devpost should contain the following:

- **Title:** Grammatical Error Detection for the Chinese Language
- **Who:** Names and logins of all your group members.
    - Rebekah Rest missing
    - Ricky Chau rchau1
    - Tanadol Lamlertprasertkul tlamlert
- **Introduction:** What problem are you trying to solve and why?
    - The paper we are implementing has the goal of detecting when and where grammatical errors occur in Chinese text.
    - This is a **classification** problem. Not only are we trying to identify the existence of grammatical errors, but we are also trying to identify the location and the kind of grammatical errors that arise within text.
- **Related Work:** Are you aware of any, or is there any prior work that you drew on to do your project?
    - Please read and briefly summarize (no more than one paragraph) at least one paper/article/blog relevant to your topic beyond the paper you are re-implementing/novel idea you are researching.
    - In this section, also include URLs to any public implementations you find of the paper you're trying to implement. Please keep this as a "living list"--if you stumble across a new implementation later down the line, add it to this list.[1]
- **Data:** What data are you using (if any)?
    - We will be using data from NLP TEA datasets (NLP TEA 2014, 2015, and 2016) which contain in total 28203 training examples and 9289 testing examples. These sentences are taken from the writing sections of TOCFL and HSK, both of which are Chinese fluency tests for foreign speakers.
    - Each training example consists of a sentence containing at least one grammatical error, the correct sentence, and a list of grammatical errors in the original sentence and their corresponding corrections. Each error can be categorized into 4 types: redundant, missing, word selection, and word ordering.

1

- The testing datasets contain both grammatical and ungrammatical sentences.

Methodology: What is the architecture of your model?

- We are training our model using the NLPTEA datasets, which contain inputs and labels for the Chinese sentences and grammatical errors.
- Our model will consist of a Recurrent Neural Network (RNN) that utilizes Long Short-Term Memory (LSTM).
- Classification will consist of 4 types of grammatical error (see chart below):

Grammatical error types	Examples of erroneous sentences	Examples of correct sentences
漏字錯誤(Missing)	我送你家 (I take you home.)	我送你**回**家
冗詞錯誤(Redundant)	他是我**的**最重要的朋友 (He is my important friend.)	他是我最重要的朋友
詞彙誤用(Selection)	我是騎腳踏車的**拿手**	我是騎腳踏車的**好手**
語序運用不當(Disorder)	我**去學校早上** (I go to school in the morning.)	我**早上去學校**

-
- If you are implementing an existing paper, detail what you think will be the hardest part about implementing the model here.

- **Metrics:** What constitutes "success?"
  - What experiments do you plan to run?
  - For this project, because we have labels and a paper to compare results to, our goal is indeed achieving an appropriate level of accuracy in our classification. The paper achieves around a 97% accuracy rating, so our goal is to be in that ballpark (>80%)..
  - If you are implementing an existing project, detail what the authors of that paper were hoping to find and how they quantified the results of their model.
  - Our **base goal** is to get a working model that successfully runs, trains, and tests
  - Our **target goal** is to get a model that can achieve a reasonable accuracy level that approaches that 97% rate found in the paper

    - Our **stretch goal** is to build a relatively efficient model that can match (or nearly match) the 97% success rate found in the paper
- **Ethics:** Choose 2 of the following bullet points to discuss; not all questions will be relevant to all projects so try to pick questions where there's interesting engagement with your project. (Remember that there's not necessarily an ethical/unethical binary; rather, we want to encourage you to think critically about your problem setup.)
    - What broader societal issues are relevant to your chosen problem space?
    - Why is Deep Learning a good approach to this problem?
    - What is your dataset? Are there any concerns about how it was collected, or labeled? Is it representative? What kind of underlying historical or societal biases might it contain?
    - Who are the major "stakeholders" in this problem, and what are the consequences of mistakes made by your algorithm?
    - How are you planning to quantify or measure error or success? What implications does your quantification have?
    - Add your own: if there is an issue about your algorithm you would like to discuss or explain further, feel free to do so.
- **Division of labor:** Briefly outline who will be responsible for which part(s) of the projects
    - We plan to all jointly partake in each step of the project so that we may each learn from all the major parts of the assignment
    - However, we may divide the major parts to have leads on each section responsible for the final product
        - Ricky: poster presentation
        - Tanadol: ____
        - Rebekah: ____