

What is a file?

- String of characters that defines a program in some format
- Can interpret these characters differently depending on context - e.g., C++ vs. Python (plaintext), vs. the a MS Word file or an image (has some encoding)
- Almost all programming languages (all mainstream ones at least) use plaintext, or plain characters, to describe their programs
- Because of this, we can read our program code as a file directly, just like any other file in the filesystem

Overview: Lexing

- Turn the program into a list of "lexemes", or distinct pieces of text, based on their character composition
- Typically splits on whitespace (outside of, e.g., comments or strings)
- Example: "(1 + 2) * 3" could turn into [LPAREN, DIGIT("1"), OP("+"), DIGIT("2"), RPAREN, OP("*"), DIGIT("3")]
- Preprocessing step that serves as a starting point for parsing. Ideally doesn't have knowledge of programming language semantics, just syntax
- Typically uses regular expressions :)